UNITED STATES AIR FORCE

SUMMER RESEARCH PROGRAM -- 1996

SUMMER RESEARCH EXTENSION PROGRAM FINAL REPORTS

VOLUME 3
ROME LABORATORY

RESEARCH & DEVELOPMENT LABORATORIES

5800 Uplander Way

Culver City, CA 90230-6608

Program Director, RDL
Gary Moore

Program Manager, AFOSR
Major Linda Steel-Goodwin

Program Manager, RDL
Scott Licoscos

Program Administrator, RDL
Johnetta Thompson

Program Administrator
Rebecca Kelly-Clemmons

Submitted to:

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Bolling Air Force Base

Washington, D.C.

December 1996

REPORT DOCUMENTATION PAGE

AFRL-SR-BL-TR-00-

0704

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	December, 1996	

4. TITLE AND SUBTITLE
1996 Summer Research Program (SRP), Summer Research Extension Program (SREP), Final Report, Volume 3, Rome Laboratory

5. FUNDING NUMBERS
F49620-93-C-0063

6. AUTHOR(S)
Gary Moore

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Research & Development Laboratories (RDL)
5800 Uplander Way
Culver City, CA 90230-6608

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Office of Scientific Research (AFOSR)
801 N. Randolph St.
Arlington, VA 22203-1977

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
Approved for Public Release

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The United States Air Force Summer Research Program (SRP) is designed to introduce university, college, and technical institute faculty members to Air Force research. This is accomplished by the faculty members, graduate students, and high school students being selected on a nationally advertised competitive basis during the summer intersession period to perform research at Air Force Research Laboratory (AFRL) Technical Directorates and Air Force Air Logistics Centers (ALC). AFOSR also offers its research associates (faculty only) an opportunity, under the Summer Research Extension Program (SREP), to continue their AFOSR-sponsored research at their home institutions through the award of research grants. This volume consists of a listing of the participants for the SREP and the technical report from each participant working at the AF Rome Laboratory.

14. SUBJECT TERMS
Air Force Research, Air Force, Engineering, Laboratories, Reports, Summer, Universities, Faculty, Graduate Student, High School Student

15. NUMBER OF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

GENERAL INSTRUCTIONS FOR COMPLETING SF 298

The Report Documentation Page (RDP) is used in announcing and cataloging reports. It is important that this information be consistent with the rest of the report, particularly the cover and title page. Instructions for filling in each block of the form follow. It is important to ***stay within the lines*** to meet ***optical scanning requirements.***

Block 1. Agency Use Only *(Leave blank)*.

Block 2. Report Date. Full publication date including day, month, and year, if available (e.g. 1 Jan 88). Must cite at least the year.

Block 3. Type of Report and Dates Covered. State whether report is interim, final, etc. If applicable, enter inclusive report dates (e.g. 10 Jun 87 - 30 Jun 88).

Block 4. Title and Subtitle. A title is taken from the part of the report that provides the most meaningful and complete information. When a report is prepared in more than one volume, repeat the primary title, add volume number, and include subtitle for the specific volume. On classified documents enter the title classification in parentheses.

Block 5. Funding Numbers. To include contract and grant numbers; may include program element number(s), project number(s), task number(s), and work unit number(s). Use the following labels:

C - Contract	**PR** - Project
G - Grant	**TA** - Task
PE - Program Element	**WU** - Work Unit Accession No.

Block 6. Author(s). Name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. If editor or compiler, this should follow the name(s).

Block 7. Performing Organization Name(s) and Address(es). Self-explanatory.

Block 8. Performing Organization Report Number. Enter the unique alphanumeric report number(s) assigned by the organization performing the report.

Block 9. Sponsoring/Monitoring Agency Name(s) and Address(es). Self-explanatory.

Block 10. Sponsoring/Monitoring Agency Report Number. *(If known)*

Block 11. Supplementary Notes. Enter information not included elsewhere such as: Prepared in cooperation with....; Trans. of....; To be published in.... When a report is revised, include a statement whether the new report supersedes or supplements the older report.

Block 12a. Distribution/Availability Statement. Denotes public availability or limitations. Cite any availability to the public. Enter additional limitations or special markings in all capitals (e.g. NOFORN, REL, ITAR).

DOD -	See DoDD 5230.24, "Distribution Statements on Technical Documents."
DOE -	See authorities.
NASA -	See Handbook NHB 2200.2.
NTIS -	Leave blank.

Block 12b. Distribution Code.

DOD -	Leave blank.
DOE -	Enter DOE distribution categories from the Standard Distribution for Unclassified Scientific and Technical Reports.
	Leave blank.
NASA -	Leave blank.
NTIS -	

Block 13. Abstract. Include a brief *(Maximum 200 words)* factual summary of the most significant information contained in the report.

Block 14. Subject Terms. Keywords or phrases identifying major subjects in the report.

Block 15. Number of Pages. Enter the total number of pages.

Block 16. Price Code. Enter appropriate price code *(NTIS only)*.

Blocks 17. - 19. Security Classifications. Self- explanatory. Enter U.S. Security Classification in accordance with U.S. Security Regulations (i.e., UNCLASSIFIED). If form contains classified information, stamp classification on the top and bottom of the page.

Block 20. Limitation of Abstract. This block must be completed to assign a limitation to the abstract. Enter either UL (unlimited) or SAR (same as report). An entry in this block is necessary if the abstract is to be limited. If blank, the abstract is assumed to be unlimited.

PREFACE

This volume is part of a five-volume set that summarizes the research of participants in the 1996 AFOSR Summer Research Extension Program (SREP.) The current volume, Volume 1 of 5, presents the final reports of SREP participants at Armstrong Laboratory. Volume 1 also includes the Management Report.

Reports presented in this volume are arranged alphabetically by author and are numbered consecutively – e.g., 1-1, 1-2, 1-3; 2-1, 2-2, 2-3, with each series of reports preceded by a 35 page management summary. Reports in the five-volume set are organized as follows:

VOLUME	TITLE
1	Armstrong Laboratory
2	Phillips Laboratory
3	Rome Laboratory
4A	Wright Laboratory
4B	Wright Laboratory
5	Arnold Engineering Development Center Air Logistics Centers

1996 SREP FINAL REPORTS

Armstrong Laboratory

VOLUME 1

1996 SREP FINAL REPORTS

Armstrong Laboratory

VOLUME 1 (cont.)

1996 SREP FINAL REPORTS

Phillips Laboratory

VOLUME 2

1996 SREP FINAL REPORTS

Phillips Laboratory

VOLUME 2 (cont.)

1996 SREP FINAL REPORTS

Rome Laboratory

VOLUME 3

1996 SREP FINAL REPORTS

Wright Laboratory

VOLUME 4A

1996 SREP FINAL REPORTS

Wright Laboratory

VOLUME 4A (cont.)

1996 SREP FINAL REPORTS

Wright Laboratory

VOLUME 4B

1996 SREP FINAL REPORTS

Wright Laboratory

VOLUME 4B (cont.)

1996 SREP FINAL REPORTS

VOLUME 5

PERFORMANCE ANALYSIS OF AN ATM-SATELLITE SYSTEM

Valentine A. Aalo
Electrical Engineering Department
Florida Atlantic University
777 Glades Road
Boca Raton, Florida 33431

Final Report for:
Summer Research Extension Program

Sponsored by:
Airforce Office of Scientific Research
Bolling Air Force Base, Washington D.C

and

Florida Atlantic University
Boca Raton, Florida 33431

December 1996

PERFORMANCE ANALYSIS OF AN ATM-SATELLITE SYSTEM

Valentine A. Aalo
Associate Professor
Electrical Engineering Department
Florida Atlantic University

Abstract

We have studied the impact of the bit error characterictics of a satellite channel on an ATM-satellite network performance. By considering the ATM header error control (HEC) mechanism, we calculate a number of ATM quality of service (QoS) parameters such as cell loss ratio and cell error rate, for the satellite link. The analysis considers both random as well as burst error channels. The results show that using additional Reed-Solomon coding to protect the ATM cells reduces cell loss rate considerable. This improved performace is at the expence of a 17% reduction in transmission rate.

PERFORMANCE ANALYSIS OF AN ATM-SATELLITE SYSTEM

Valentine A. Aalo

1. Introduction

The switching and multiplexing technique defined for Broadband Integrated Services Digital Network (B-ISDN) is the Asynchronous Transfer Mode (ATM), and is designed to provide a single common format for transporting voice, video, and data in an integrated network. ATM requires that the information from one or several users be multiplexed, buffered, and partitioned into blocks of fixed length, called cells. The fixed-length cell structure allows ATM to support applications with continuous bit-rate and variable bit-rate simultaneously. Each ATM cell consists of 5 bytes of header and 48 bytes of payload. The header contains information about channel identification, priority, payload type, and error control mechanism. The payload carries the actual information to be transmitted. The mode of information transfer is connection-oriented and as such, a virtual connection must be established between the source and the destination before the information transfer can take place. ATM is independent of the information transmission system provided and may be transported using any digital transmission format [1].

ATM was designed for transmission over a physical channel with excellent error characteristics, such as the fiber optic channel where the error rate is very low and due mostly to random errors. In a wireless environment such as a satellite channel, however, the error rate is usually very high, and bit errors usually occur in bursts due to multipath, interference, and fading. Spectrum congestion resulting from increased demand for wideband telecommunication services has forced satellite designers to consider operating at very high frequencies, where atmospheric propagation effects, especially attenuation and fading due to rain, are very severe. For example, NASA's Advanced Communications Technology Satellite (ACTS) operates in the Ka-band (20/30 GHz) where rain attenuation effects are particularly severe. Also, satellite systems are power limited and usually require the use of forward error correction (FEC) to improve the quality of the transmission. Residual errors associated with the decoders used extensively with the receiver

on satellite links also tend to be bursty. Today, due to their ability to offer wide bandwidth on demand and allow flexibility regarding topology, reconfiguration, or network expansion, satellites are used extensively in wideband communication systems. Therefore, it is expected that satellites will play an important role in the provision of ATM services. For example, ATM transmission at OC-3 (155.54 Mb/s) has been successfully demonstrated over the ACTS, and commercial ATM service is available at 45Mb/s [2]. However, many performance issues regarding the reliable transmission of ATM signals over a satellite channel remain to be resolved. In this study we focus on the impact of the bit error characteristics of a satellite channel on the ATM layer.

This report is organized as follows. Section 2 discusses tha satellite channel error characterization including the type of modulation and channel coding used. In section 3, the specific ATM quality of service (QoS) parameters under consideration are defined and quantified. Then the impact of both random and bursty channel errors on the QoS parameters is then analyzed. Some techniques to improve the ATM system performance are discussed in section 5 while some concluding remarks are given in section 7.

2. Satellite Channel Characterization

The performance of an ATM-Satellite network depends largely on the bit error characteristics of the underlying physical link. In the fiber optic link for which ATM was originally designed the bit errors are randomly distributed and very low. However, in a wireless environment and in a satellite channel in particular, bit error rates are usually high and the errors are sometimes bursty. A geostationary satellite channel can be approximated as an Additive White Gaussian Noise (AWGN) channel with random errors. Because the satellite system is power-limited, M-ary phase shift keying (MPSK) modulation will be considered. In this case, assuming perfect phase coherence, the bit error rate (BER) for an uncoded system is well approximated by

$$p_e \approx 2Q\left(\sqrt{\frac{2E_s}{N_0}}\sin(\frac{\pi}{M})\right) \tag{2.1}$$

where

$$E_s = (\log_2 M)E_b,$$

$$Q(y) = \frac{1}{\sqrt{2\pi}}\int_y^\infty e^{-x^2/2}dx,$$

and $\frac{E_b}{N_0}$ is the bit energy to noise density ratio. Since power at the satellite is limited, to improve the performance of satellite communication systems forward error control coding is usually employed. On an average, coding reduces the bit error rate but requires the addition of redundancy in the transmission. Viterbi decoders are usually used at the receiver in satellite systems and results in the occurrence of burst errors. Typically, the coding system use concatenated codes with inner convolutional codes with soft-decision Viterbi decoder combined with outer Reed-Solomon codes. In such case, when the Viterbi decoder strays erroneously onto an incorrect trellis path, it results in a burst of errors at the decoder output. Since ATM was designed to be robust to random single bit errors, the presence of burst errors considerably degrade the performance of the ATM-satellite system. Thus the outer code is designed to correct the burst error provided that the burst length does not exceed the error correcting capability of the bounded distance of the outer code.

A commonly used coding scheme in satellite modems is the Reed-Solomon code which achieves the largest possible code minimum distance for any linear code with the same encoder input and output lengths and therefore, are particularly useful for burst-error correction. The Reed-Solomon decoded symbol error probability may be expressed in terms of the channel symbol error, p, as [3]

$$p_e \approx \frac{1}{n} \sum_{j=t+1}^{n} j \binom{n}{j} p^j (1-p)^{n-j} \tag{2.2}$$

where n is the Reed-Solomon block symbol code length, k is the number of data symbols being encoded, and

$$t = \frac{n-k}{2}.$$

Thus in the MPSK system under consideration, p is given in (2.1) with E_s replaced by $\left(\frac{k}{n}\right) E_s$. When the effect of Reed-Solomon coding is considered it will be assumed that the entire ATM cell is protected at a time, that is, $k = 53 \times 8 = 424$ bits and $n = 511$ bits.

3. ATM QoS Parameters

A number of quality of service parameters have been identified as being very important in assessing the performance of the ATM protocol. Recall that the ATM cell of 53 bytes comprises of 5 bytes of header and 48 bytes of payload.

The header is again subdivided into 4 bytes for the information field and 8 bits (1 byte) for header error control (HEC). The HEC which is used to protect the header information from transmission errors can detect certain multiple bit errors but can correct only single bit errors. ATM was designed to be robust to random single bit errors. During the transmission, the ATM header may have no errors, a single bit error, or multiple bit errors. If single bit errors occur in the header, it is usually detected and corrected by the HEC mechanism. Since the HEC can detect both single bit errors and multiple errors but can correct only single bit errors, two events are possible when multiple bits in the header arrive in error. If the errors are detected, the entire ATM cell is discarded. Otherwise, if the errored bits are undetected, the ATM cells are transmitted with false valid address leading to the misinsertion of the cell into a wrong address. Although errors may occur in the payload, the ATM payload has no error correction mechanism. Fig. 1 summarizes the ATM error characteristics.

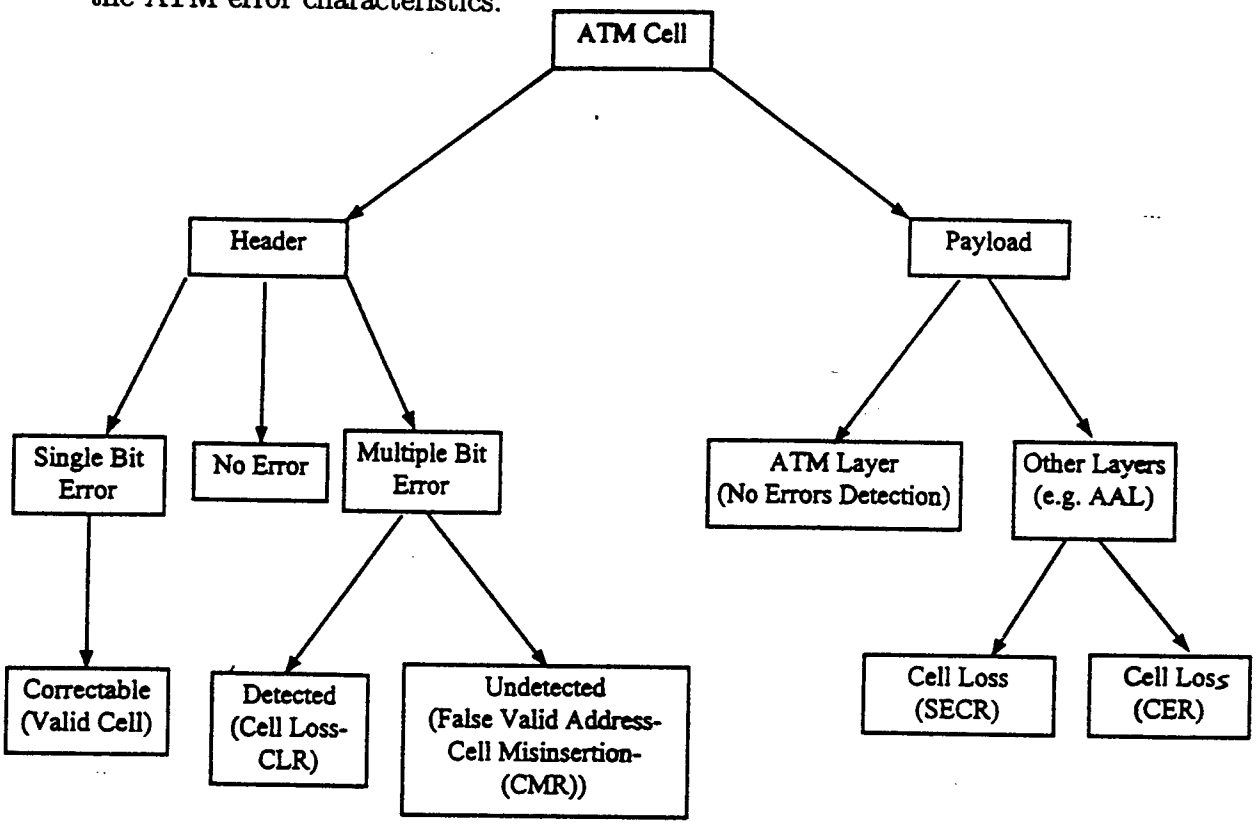


Figure 1: ATM error characterization

Depending on the nature of the error and whether error is detected and corrected, the following QoS parameters will be considered.

Cell Loss Ratio (CLR):

This is the ratio of the number of lost ATM cells to the total number of cells sent by the user in a specified time. Due to the statistical nature of the ATM multiplexing, the limited size of the ATM buffer, the complex flow control mechanism, and the random nature of the transmission channel, it is usually possible that the header HEC is unable to correct the errored bits leading to ATM cells being lost. On the average, the probability with which ATM cells are lost in the system is the cell loss ratio. It is given by the probability that two or more header bits arrive in error.

In practice, in order to reduce the vulnerability of the system to burst errors a two-state Markov model [4], shown in Fig.2, may be used. Under normal conditions, the receiver is in the single bit error correction mode and remains in this mode as long as no errors are detected. However, if a single bit error is detected, it is corrected and the receiver switches to the detection mode. If multiple bits are detected, since more than one bit cannot be corrected by the HEC, the ATM cell is discarded and the receiver makes a transition to the detection mode. When the receiver is in the detection mode, no attempt is made to correct errors, and all errored cells (including single bit errors) are discarded. The receiver remains in this mode as long as the cells are received with error. When the header is examined and found to be error free, the receiver makes a transition to the correction mode. Thus with the Markov model, a cell is discarded if the receiver is in the correction mode and more than one bit error occurs or if the receiver is in the detection mode and at least one bit error occurs in the header [5].

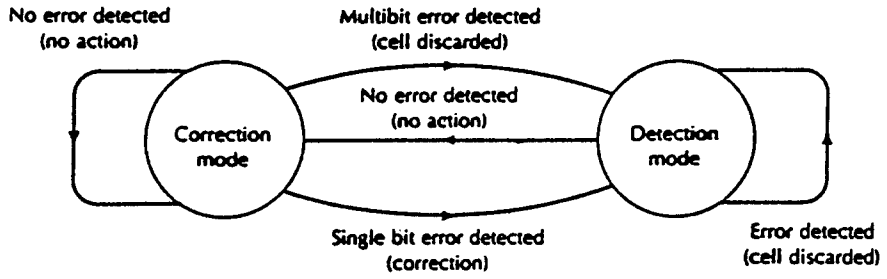


Figure 2: Two-state Markov model for HEC operation

Cell error ratio (CER):
This is the ratio of errored ATM cells (that is, cells with at least one error in the payload) to the total number of successfully delivered cells.

Severely Errored Cell Ratio (SECR):
This is defined as the ratio of severely errored ATM cells to the total number of successfully delivered cells. A cell is referred to as severely errored if N ($N \geq 2$) errors occur in the payload of the ATM cell. Throughout this report we assume $N = 2$.

Cell Misinsertion Ratio (CMR):
This is defined as the ratio of the cells delivered to a wrong destination to the total number of cells sent. A cell misinsertion occurs as a result of an undetected error in the header that causes a change of the cell destination resulting in the misinsertion of the cell a wrong virtual channel or virtual path.

4. Effect of Channel Errors on ATM Parameters

In this section we consider the effect of the satellite channel error characteristics on the ATM quality of service parameters. The effect of both random errors (typical of optical fiber link) and burst errors (typical on a satellite link) are considered. For the uncoded system, the probability of channel bit error, p_e, given in (2.1) will be used. However, for the case that additional Reed-Solomon coding is used to improve system performance, (2.2) will be used for the channel bit error probability.

4.1. Random Errors

We assume that the occurrences of single random bit errors in the channel be independent and identically distributed, with probability p_e. Then the probability that k out of the 40 header bits arrive in error is binomially distributed and is given by

$$p_s(k) = \binom{40}{k} p_e^k (1 - p_e)^{40-k} \tag{4.1}$$

Then it can be shown that if the receiver operates with the dual-mode Markov model, the cell loss ratio P_{CLR} is given by [5]

$$P_{CLR} = p_c\,[1 - p_s(0) - p_s(1)] + (1 - p_c)\,[1 - p_s(0)] \tag{4.2}$$

where

$$p_c = p_s(0) = (1 - p_e)^{40}.$$

To obtain the cell error ratio P_{CER}, we compute the probability that one or more error bits occur in the payload of the ATM cell. Since there are $48 \times 8 = 384$ bits in the information field, we have

$$P_{CER} = 1 - (1 - p_e)^{384} \tag{4.3}$$

In the case of severely errored cell ratio P_{SECR}, the probability that at least two bit errors occur in the information field is given by

$$P_{SECR} = 1 - (1 - p_e)^{384} - 384 p_e (1 - p_e)^{383}. \tag{4.4}$$

4.2. Burst Errors

Typically, satellite links are susceptible to the occurrence of burst errors resulting not only from the dynamic nature of the channel but also from coding schemes that are commonly used to improve the power efficiency of the satellite link. Thus multiple bit errors occur frequently. One commonly used model to characterize the bursty nature of the channel assumes that the error bursts as well as errors in the burst are Poisson distributed leading to the Neyman-A contagious model [5]. In this model, the probability that k errors occur in an interval of m bits is given by [5]

$$p_b(k) = \frac{b^k}{k!} \exp\left(-\frac{mp_e}{b}\right) \sum_{j=0}^{\infty} \left(\frac{mp_e}{b} e^{-b}\right)^j \frac{j^k}{j!} \tag{4.5}$$

where p_e is the channel bit error rate and b ($6 \le b \le 40$) is the mean error burst length. Throughout we use $b = 6$ to obtain the numerical results. It can be shown from (4.5) that

$$p_b(0) = \exp\left[-\frac{mp_e}{b}\left(1 - e^{-b}\right)\right] \tag{4.6}$$

and

$$p_b(1) = (mp_e) \exp\left[-\frac{mp_e}{b}\left(1 - e^{-b}\right) - b\right]. \tag{4.7}$$

Therefore, the cell loss ratio for the bursty satellite link using the two-state Markov model is given by

$$P_{CLR} = p_c\left[1 - p_b(0) - p_b(1)\right] + (1 - p_c)\left[1 - p_b(0)\right]. \tag{4.8}$$

Also, the cell error ratio in this case is given by

$$P_{CER} = 1 - \exp\left[-\frac{384p_e}{b}\left(1 - e^{-b}\right)\right] \tag{4.9}$$

and the severely errored cell ratio is given by

$$P_{SECR} = 1 - \exp\left[-\frac{mp_e}{b}\left(1 - e^{-b}\right)\right] - (384p_e)\exp\left[-\frac{mp_e}{b}\left(1 - e^{-b}\right) - b\right]. \tag{4.10}$$

5. ATM Performance Enhancement

ATM cells are usually transmitted over fiber optic channels where the error rate is not only very low but the bit errors are also randomly distributed. However, when ATM cells are transmitted over a satellite channel where error rates are high and the bit errors occur in bursts, the system performance is expected to degrade. This is because in the ATM-satellite link with bursty errors, these errors cannot be corrected since the ATM header is capable of correcting only single errors. In such a system there are a number of techniques that can be used to improve system performance. Some techniques that will be discussed in this section are; cell loss recovery using FEC, interleaving to spread the error bursts into dispersed single-bit errors, HEC extension to correct multibit errors, and use of additional coding for tha ATM cell stream.

5.1. Cell Loss Recovery

Typically, two major factors cause ATM networks to discard cells. One factor is bit errors in the physical channel. As stated earlier, since the HEC can correct single bit errors in the header field, cell loss due to uncorrectable header error in a random bit environment reduces greatly. If the bit error rate is small ($p_e \leq 10^{-9}$) as in an optical transmission channel, the probability of cell discard due to bit errors may be neglected. Burst errors, on the other hand, encountered in a wireless network such as the satellite channel, can be a major source of cell delineation. Although cell loss may not occur too frequently, once it occurs, consecutive cells are also discarded. Therefore, at least 13 cells are usually consecutively discarded

whenever a cell loss occurs. The other cause of cell loss is due to buffer overflow in multiplexing. When buffer overflow occurs in an ATM node, consecutive overflows may usually be expected because the buffer is close to full. A two-state Markov model may also be used to model the cell discard process due to buffer overflow [6].

A number of cell loss recovery techniques based on forward error control have been used to improve the performance of ATM networks. These range from no protection of the ATM cells to using a variety of error detection/correction techniques. Thus, while some of these techniques deal with only single cell loss, others deal with consecutive cell loss compensation. A commonly used technique consists of cell loss detection and lost cell regeneration and is applicable to the virtual paths of the ATM network. However, this technique is more effective in optical networks where transmission errors are low [6].

5.2. Interleaving

An FEC based cell recovery technique may not be desirable in a heavily wireless network since, in general, the FEC method requires the transmission of redundant cells, thus increasing the total network traffic. On the other hand, the method of interleaving has a very small overhead and can improve the cell loss and insertion probabilities in systems that are vulnerable to burst errors without requiring complex interfacing. Two approaches may be used to apply interleaving to ATM cells. One approach is to interleave the headers of several cells in a block of ATM cells (also called block unit interleaving or inter-cell interleaving). A particular block unit interleaver (the ATM Link Enhancer) reduces cell loss probability by performing the interleaving of headers of ATM cells in a unit block of 12 ATM cells and has been shown to provide considerable performance improvement on a DS3 ATM-satellite link [7]. Another approach, known as cell unit interleaving (also called intra-cell interleaving), disperses each bit of the header over the entire data field of the particular ATM cell. Although not as effective as block unit interleaving, this approach has a much smaller processing delay and is more effective when burst lengths are small (shorter than 11 bits) [8].

5.3. HEC Extension

The ATM header error control scheme may detect multiple bit errors but is capable of correcting only single errors. In the HEC extension method, the error control scheme typically used in the ATM header is replaced with a more powerful code

that is capable of correcting multiple bit errors. The residual bit error rate may be reduced considerably. However, this method is usually incompatible with the standard ATM header format and may require complex interfacing [8].

5.4. Additional Coding

The use of block interleaving requires a continuous stream of ATM cells which may not be readily available in very small aperture terminal (VSAT) systems. Although cell unit interleaving may used in this case, however, the associated performance gain may be too small to be effective when the burst lengths are long. Better performance improvement may be expected if additional coding is used on the satellite link to protect the ATM cells. This introduces additional overhead and thus reduces the data rate. In this work, we consider the effect of using Reed-Solomon coding to protect each ATM cell at a time.

6. Numerical Results

In this study we examine the effect of channel bit error on a number of ATM QoS parameters. We consider both random errors and burst errors (based on equation (4.5) with $b = 6$). For both random and burst errors, we first assume that no additional coding is provided for the ATM cell protection. We refer to this as the *uncoded* performance. Then we consider the simplified case where each ATM cell (424 bits) is block coded with a $(424, 511)$ Reed-Solomon code. This is referred to as the *coded* performance. In Figs. 3 and 4, the cell loss ratio (P_{CLR}) is plotted against the channel bit error probability (p_e) for both channels with random errors and burst errors, respectively. We observe that there is considerable decrease in cell loss rate with the additional Reed-Solomon coding. Since, for a given value of p_e, the type of modulation (MPSK) and power level (E_b/N_0) will have to be specified, the rest of the results consider the effect of M and E_b/N_0 on the ATM QoS parameters. The QoS parameters for uncoded ATM cells are shown in Figs. 5-7 for a channel with random bit errors and in Figs. 8-10 for a burst error channel. The corresponding results for coded ATM cells are given in Figs. 11-13 for a channel with random bit errors and in Figs. 14-16 for a burst error channel.

7. Conclusion

In this report we have studied the effect of using Reed-Solomon coding to protect the ATM cells to be transmitted over a satellite channel. The simplified link considered assumes that the channel is an AWGN channel and the modulation QPSK, 8PSK, or 16PSK. For a given modulation level M, the SNR E_b/N_0 required to obtain a given QoS performance level can be obtained for both the random bit error channel and the bursty channel. It is observed that a considerable gain in performance is obtained. However, this gain is at the cost of a 17% reduction in the transmission rate.

References

[1] R. O. Onvural, *Asynchronous Transfer Mode Networks Performance Issues*, Boston, MA: Artech House, 1994.

[2] L. P. Seidman, "Satellites for wideband access", IEEE Communications Magazine, vol. 34, pp.108-111, October 1996.

[3] B. Sklar, *Digital Communications, Fundamentals and Applications*, Englewood Cliffs, NJ: Prentice Hall, 1988.

[4] J. R. Yee and E. J. Weldon, "Evaluation of the performance of error-correcting codes on a Gilbert channel", IEEE Trans. Commun., vol. COM-43, pp.2316-2323, August 1995.

[5] S. Ramseier and T. Kaltenschnee, "ATM over satellite: analysis of ATM QoS parameters", Proceedings of ICC, pp. 1562-1566, June 1995.

[6] H. Ohta and T. Kitami, A cell loss recovery method using FEC in ATM networks", IEEE JSAC, vol. 9, pp. 1471-1482, December 1991.

[7] D. M. Chitre et al., "Asynchronous Transfer Mode (ATM) Operation via satellite: issues, challenges and resolutions", International Journal of Satellite Commun., vol.12, pp. 211-222, May/June 1994.

[8] S. H. Lim and D. M. An, "Impact of cell unit interleaving on header error control performance in wireless ATM", Proceedings of Globecom, pp. 1705-1709, November 1996.

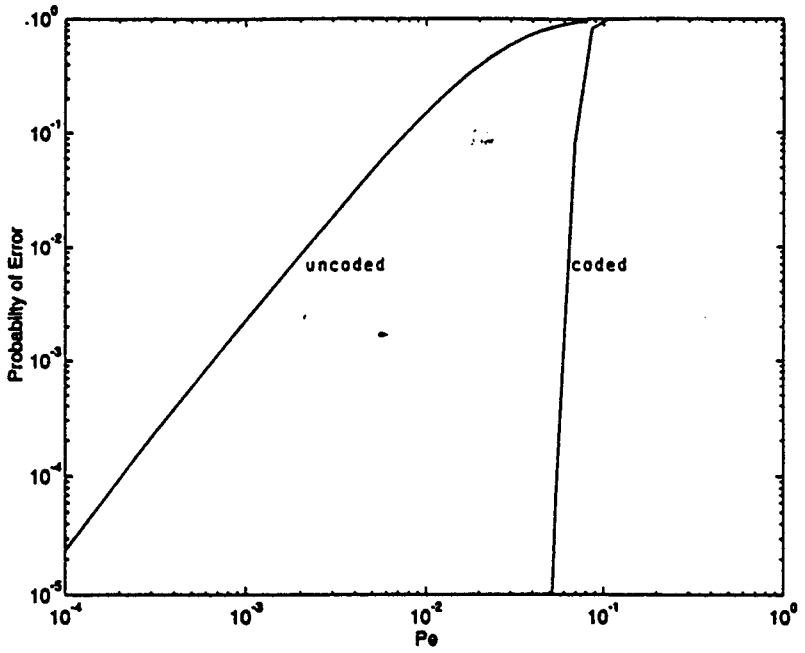


Figure 3: Cell loss ratio versus bit error rate for a random error channel

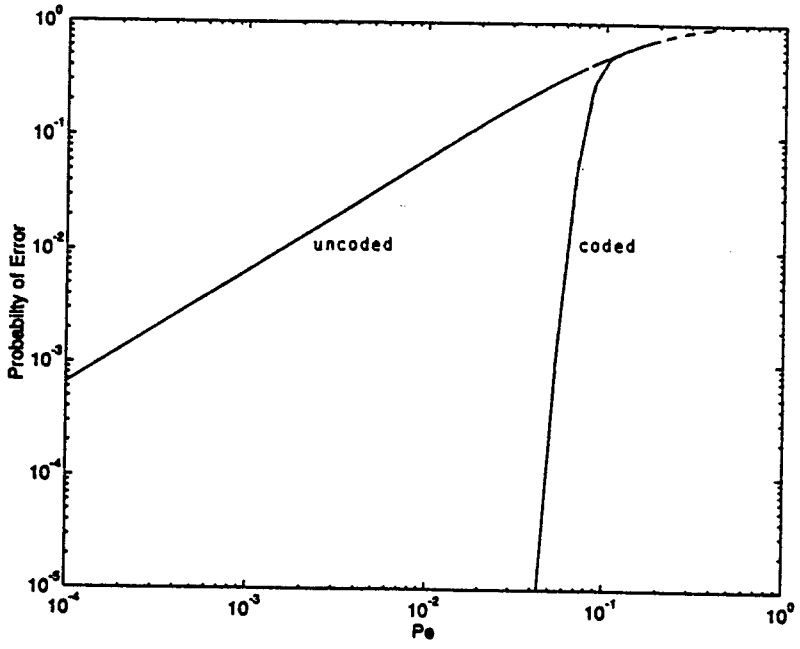


Figure 4: Cell loss ratio versus bit error rate for a burst error channel

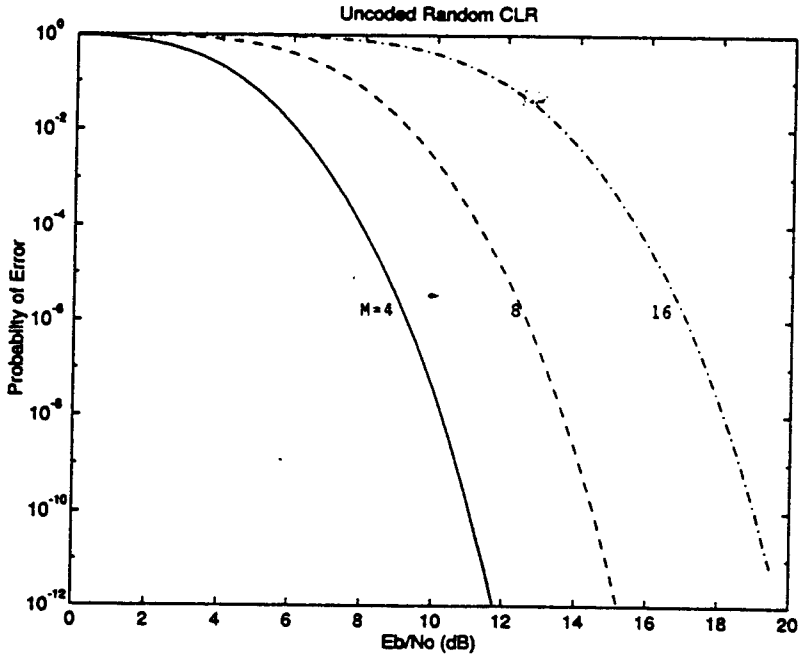


Figure 5: Cell loss ratio for uncoded ATM cells on a random bit error channel

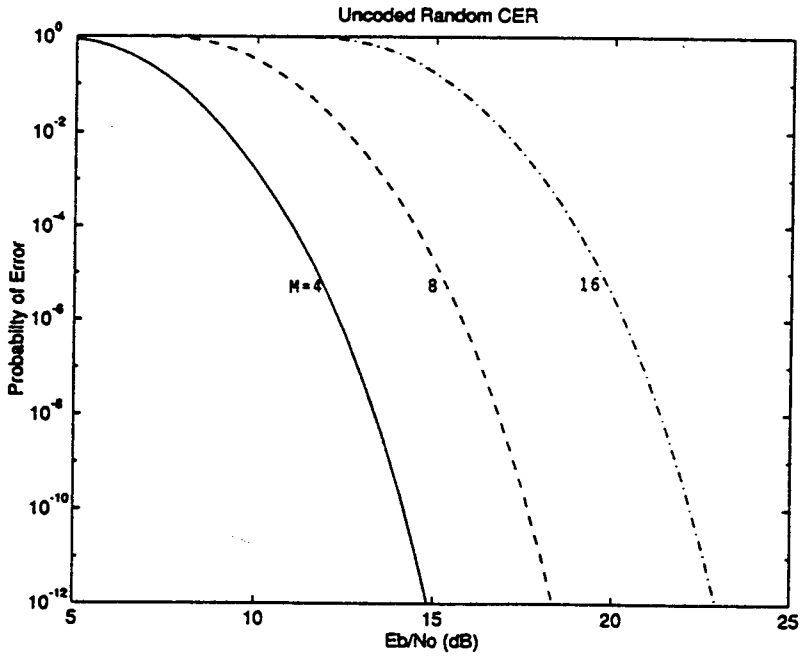


Figure 6: Cell error ratio for uncoded ATM cells on a random bit error channel

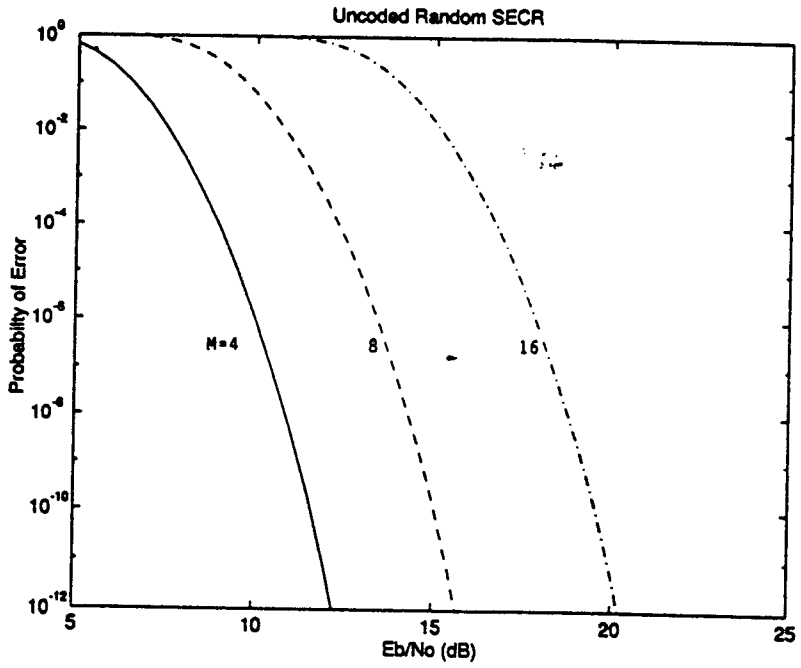


Figure 7: Severely errored cell ratio for uncoded ATM cells on a random bit error channel

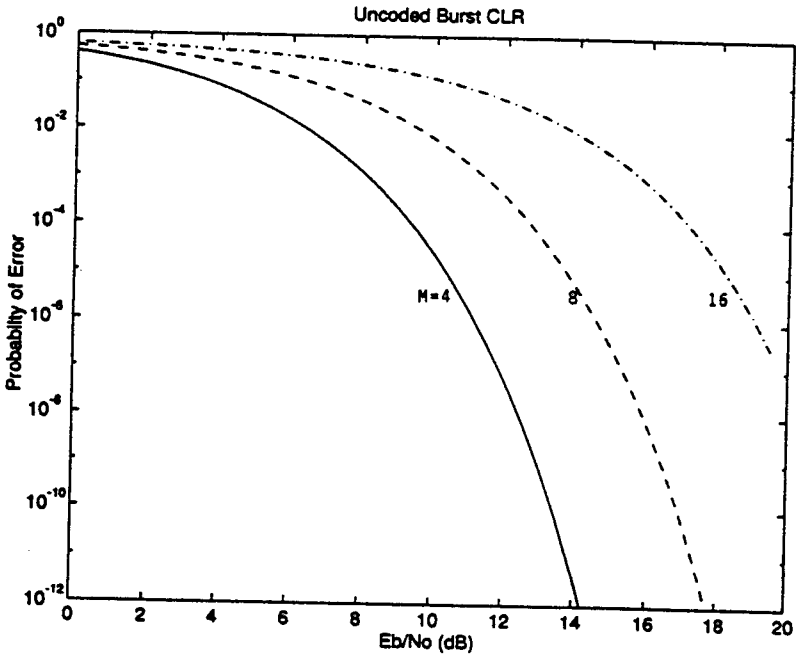


Figure 8: Cell loss ratio for uncoded ATM cells on a burst error channel

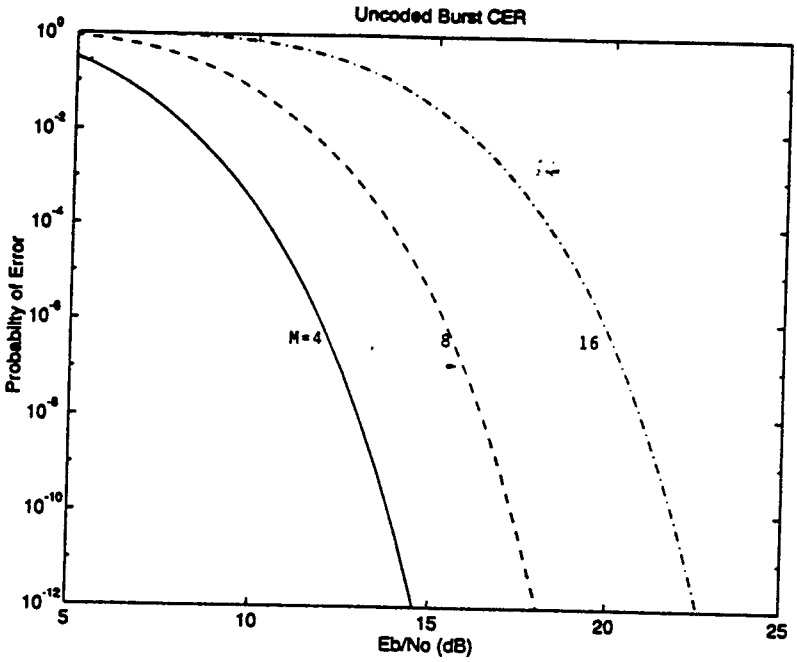


Figure 9: Cell error ratio for uncoded ATM cells on a burst error channel

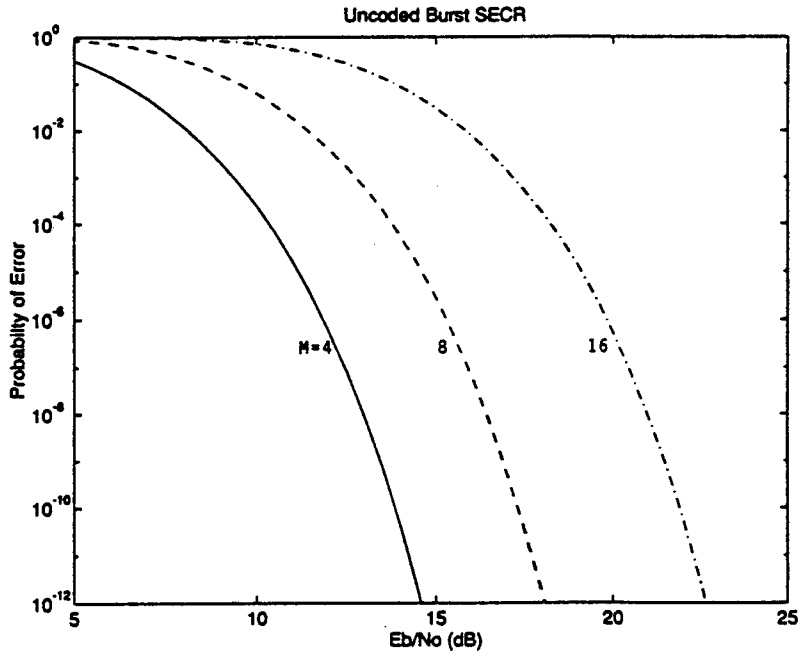


Figure 10: Severely errored cell ratio for uncoded ATM cells on a burst error channel

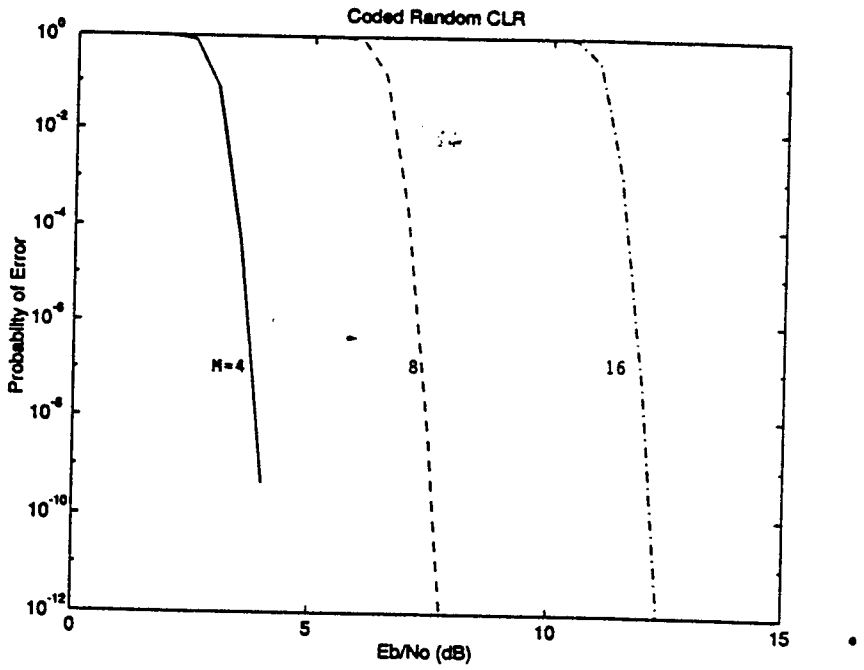


Figure 11: Cell loss ratio for coded ATM cells on a random bit error channel

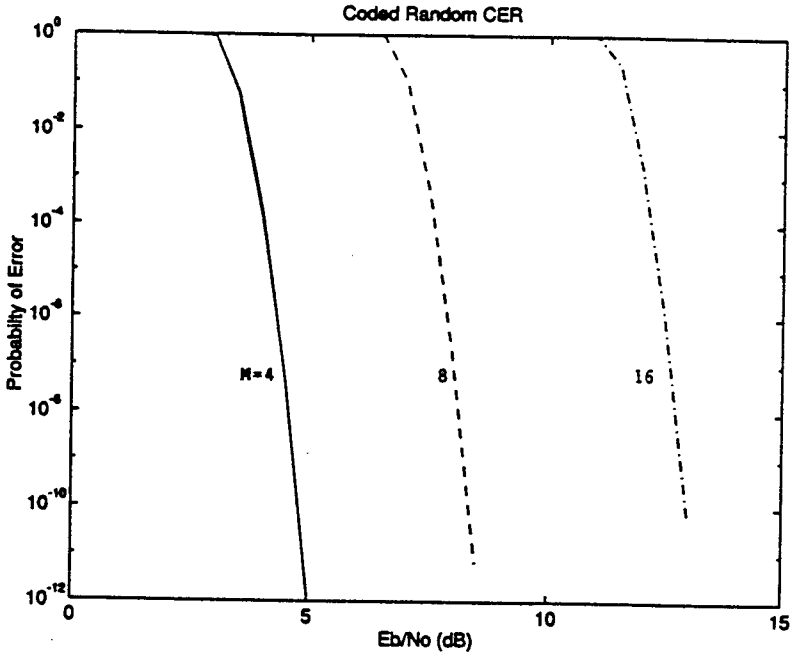


Figure 12: Cell error ratio for coded ATM cells on a random bit error channel

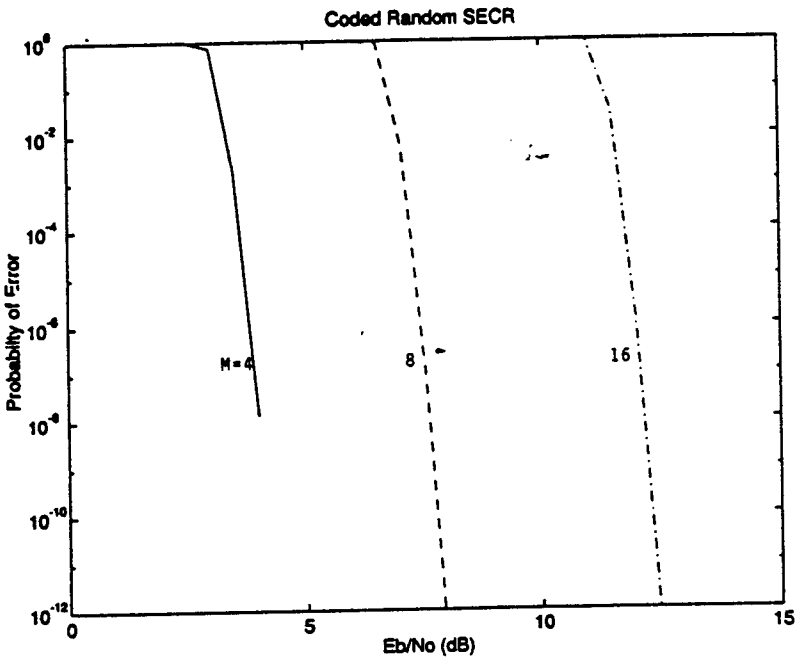


Figure 13: Severely errored cell ratio for coded ATM cells on a random bit error channel

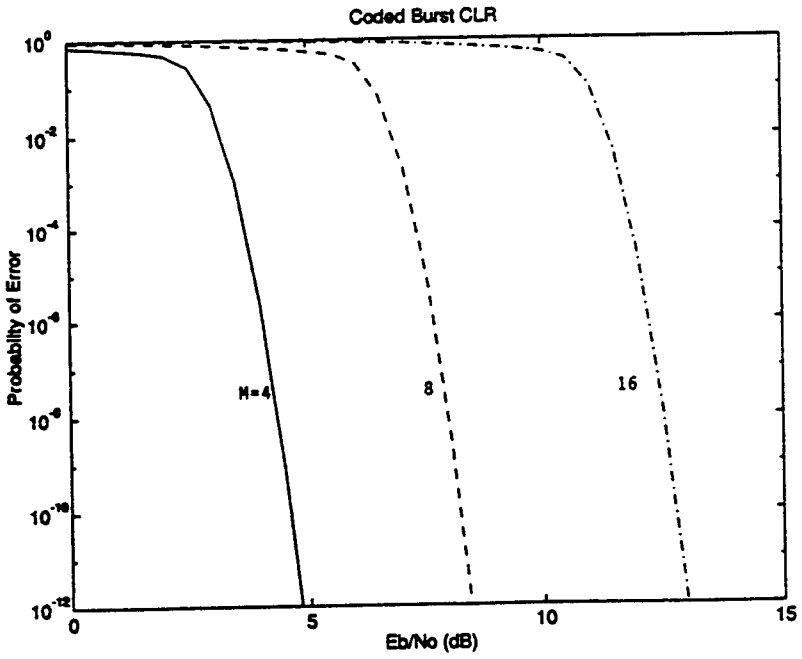


Figure 14: Cell loss ratio for coded ATM cells on a burst error channel

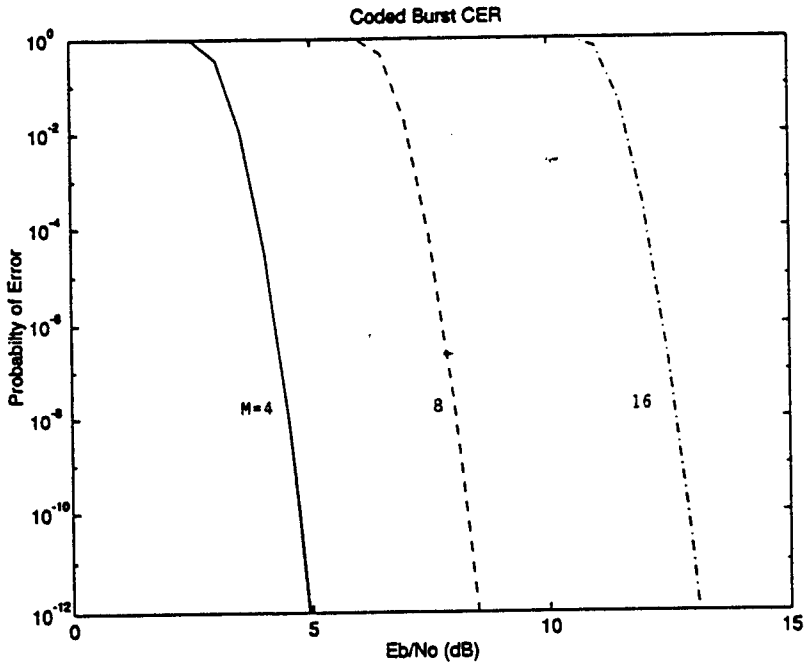


Figure 15: Cell error ratio for coded ATM cells on a burst error channel

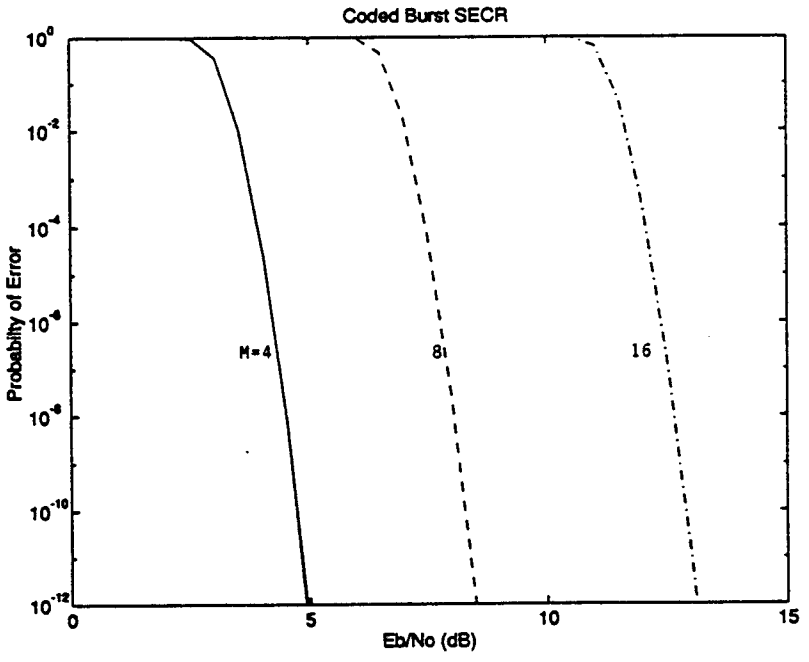


Figure 16: Severely errored cell ratio for coded ATM cells on a burst error channel

Formal Specification for Job-Shop Scheduling Problems

Blayne Mayfield
Associate Professor
Department of Computer Engineering

Harapan Sinaga
Graduate Student

Paul Benjamin
Assistant Professor

Oklahoma State University
Stillwater, OK 74078

Final Report for:
Summer Research Extension Program

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D.C.

and

Rome Laboratory

December 1996

FORMAL SPECIFICATION FOR JOB-SHOP SHCEDULING PROBLEMS

Harapan Sinaga
Paul Benjamin
Blayne Mayfield
Computer Science Department
Oklahoma State University

Abstract

Many researchers have been paying attention to formalize and automate various sources of programming knowledge and integrate them into highly automated tools for developing specifications into correct and efficient software programs. Such tools serve to raise the level of language from which programmers can obtain correct and efficient executable codes through the use of the automated tools. One such tool is KIDS -- the Kestrel Interactive Development System. It provides automated support for the development and has components for performing algorithm design, deductive inference, optimizations, data type refinement, and compilation.

Unfortunately, although this system provides logical tools that help the user to specify correct programs, it does not help the user to find efficient designs. The goals of this research are to design and implement job-shop scheduling problem theories by decomposition using KIDS and to identify and incorporate useful scheduling heuristics for such problems.

FORMAL SPECIFICATION FOR JOB-SHOP SHCEDULING PROBLEMS

Harapan Sinaga
Paul Benjamin
Blayne Mayfield

Introduction

Many researchers have been paying attention to formalize and automate various sources of programming knowledge and integrate them into highly automated tools for developing formal specifications into correct and efficient software programs. Such tools serve to raise the level of language from which programmers can obtain correct and efficient executable code through the use of the automated tools. These tools help users design and develop from formal specification methods. One such tool is KIDS – the Kestrel Interactive Development System. It provides automated support for the development and has components for performing algorithm design, deductive inference, program simplification, partial evaluation, finite differencing optimizations, data type refinement, compilation, and other development operations [Smith, 1990]. Unfortunately, although this system provides logical tools that help the user to specify correct programs, it does not help the user to find efficient designs.

This research is a continuation of the research initiated by one of the authors, Paul Benjamin. He conducted research to design software for a class of problems such as N-queen problems. In this project we focus our research to deal with a class of specific problems, namely job-shop scheduling problems. The purpose of this research is to designing and implementing job-shop scheduling problem theories by decomposition using KIDS and identifying and incorporating useful scheduling heuristics for such problems.

KIDS -- Kestrel Interactive Development System

In our research, we will use KIDS as the automated tool. KIDS is a semi-automatic program transformation system in which users apply a sequence of consistency-preserving transformations to an initial formal specification of a problem and get a correct and efficient program for the problem. It emphasizes the application of complex-high level transformations that perform significant and meaningful actions. KIDS combines a theory of problem solving with a domain theory to derive a high level program, which can be transformed to improve efficiency.

KIDS features algorithm design tactics and deductive inference components. It is capable of automated supports for the development and has components for performing algorithm designs, deductive inferences, optimizations, data type refinement, and compilation. The optimizations are program simplification, partial evaluation, finite differencing, and other development operations.

KIDS is currently run on Symbolics, SUN-4, and SPARC workstations. It is built on top of REFINE, a knowledge-base programming environment. Its environment provides an object-attributed-style database to represent software-related objects via annotated abstract syntax trees, grammar-based parser/un-parser translating between text and abstract syntax, and a very high level language and its compiler. The language supports first-order theory, set-theoretic data types and operations, transformations and pattern constructs supporting the rule creation. The compiler generates codes.

Typical steps developing a program using KIDS are the following [Smith 1990a]

- Developing a domain theory for a problem in terms of a collection of data types, operations, laws, and inference rules;
- Creating a specification for the problem in terms of the underlying domain theory;
- Performing a design tactic by selecting an algorithm and applying it the specification;

- Applying optimizations by a selecting optimizing methods and applying to an expression in the program;
- Applying data type refinements by selecting detail implementations for high level data types in the program; and
- Compiling to get an executable code.

As mentioned above, in order to use KIDS, a user builds up a domain theory. A domain theory defines the basic concept of a problem in terms of functions, data types and laws and inference rules to reasoning about the concepts. It specifies the concept, operations, and the relationship that characterize the problem and support reasoning about the domain via a deductive inference rule system. It is represented using first order theory.

A formal specification serves to define the problem for which we desire an efficient computational solution in terms of the underlying domain theory. The problem is defined by means of functional constraints on the input/output behavior. A specification will be represented as a quadruple F = <D, R, I, O> in a more program-like format:

function F(x : D) : set(R)

where I(x)

returns {z | O(x, z)}

= Body.

This quadruple means that function F needs input type D that satisfies input condition I and results output type R that satisfies output condition O. If exist, Body is a code to be executed to compute F. The input condition or input assumption is a map from D to set {true, false}:

$I : D \rightarrow$ Boolean

and the output condition is a map from D×R to set {true, false}:

$O : D \times R \rightarrow$ Boolean.

This specification for problem F returns the set of all values z of type R satisfying the output condition O. This is consistent if for all possible inputs satisfying the input condition, the body produces the same set as specified in the returns expression, formally

$$\forall(x \in D)\ (I(x) \Rightarrow F(x) = \{z \in R \mid O(x, z)\}$$

Those will be represented as axioms in the extended first order theory using λ-calculus like format.

Job-Shop Scheduling Problem Formal Specification

The essential notion of scheduling is that certain activities are assigned to resources over certain time intervals. Various constraints on the assignment must be satisfied and certain measure of the cost or goodness of the assignment is to be optimized. For job-shop scheduling problems, activities are a set of jobs, each consisting of sub-jobs or tasks, and resources are a set of machines. Jobs are to be scheduled by assigning them to machines. A time interval model is used to partition the activities. In this section, we will develop a scheduler involving several steps:

- Developing a formal model of the job-shop scheduling domain called a domain theory;
- Stating the constraints, objective, and preferences of job-shop scheduling problems within a domain theory as a problem specification; and
- Selecting and applying a design tactic and identifying heuristic filters for the problem to get efficient code.

A domain theory for a problem defines the basic concept of the problem in terms of functions and data types and laws and inference rules to reasoning about the concepts. It specifies the concept, operations, and the relationship characterizing the problem and supporting reasoning about the domain via a deductive inference rule system. It is represented using a formal language.

A domain theory for a job-shop scheduling problem defines the basic concepts of scheduling and the laws for reasoning about the concepts. The general components of a domain theory for job-shop scheduling problems are jobs – including to state their internal structures and characteristics, hierarchies, and various operations on jobs; machines -- including to state their internal structures and characteristics, hierarchies, and various operations on machines; time -- including to state a calculus of discrete time partition; constraints -- including to state the condition to be satisfied and a calculus for reasoning about them; objectives – including to state the cost of a schedule to be minimized; and a schedule.

In job-shop scheduling problems, several classes of constraints commonly arise. The most common constraints are precedence constraints and capacity constraints. The precedence constraints are to state an activity that must precede others. That is, tasks are partially ordered. Meanwhile, the capacity constraints are to state bounds on the capacities of resources. Other constraints may involve [Fox et al., 1989].

There are several objectives in job-shop scheduling problems. Typically, it is to minimize the cost of a schedule. Cost can be measured in terms of time to completion, due time, earliness, tardiness, work-in-progress, total cost of consumed resources, or resource utilization. These determine the goodness of a schedule.

Using the above concept, a job-shop scheduling problem can be formulated. A schedule to be found is a set of assignments satisfying those constraints and objectives. Some constraints that are usually in job-shop scheduling problems are:

- Consistent-Ordering

  Each job goes through requested machines for a service in certain order

- Consistent-Machine-Assignment

  No machines serve more than one job at any time.

- Consistent-Job-Assignment

  No job is assigned to more than one machine at any time.

- Consistent-Release-Time

  The start time of jobs assigned to a machine must not precede before their availability.

- Consistent-Due-Time

  If any, the finish time of jobs must be no later than their due time.

- Consistent-Scheduling

  All jobs must be scheduled.

To simplify the problems, all necessary movements of jobs from one machine into another are already counted including an operator for the machine. It is possible however to include these certain aspects of job-shop scheduling problems. Also, no jobs visit certain machines more than once and there is no preemption; i.e., if a machine services a job, the service cannot be interrupted until the job is done. In addition, we may assume that all machines are available during the whole operation and the number of machines and the number of jobs to be scheduled are fixed and known. The number of jobs is N, the number of machines is M and there is no duplicate machines.

In job-shop scheduling problems, objectives are functions used to measure the "goodness" of the schedule. Some common objective functions are:

- Minimizing-Flow-Time – The purpose of this objective function is to minimize the flow time, i.e., the sum of the completion times of the whole jobs in the system. It gives an indication of the total holding or inventory, costs incurred by the schedule. It is common to include weighted in the completion times, called a weighted flow time.
- Minimizing-Makespan – The purpose of this objective function is to minimize the completion of the last job to leave the system.

- Minimizing-Maximum-Lateness – The purpose of this objective function is to minimize the worst violation of the due times of jobs.

In addition some policies may be enforced. These could be the "longest processing time first," the "shortest processing times first," etc. However, we do not consider such policies in this research.

In general, a job-shop scheduling problem is to construct a set of assignments. These assignments are called a schedule. Formally, it is a simple relation:

Schedule : set(Job × Machine × Time)

A scheduler assigns a set of jobs into a set of machines at certain times for given constraints and good values of an objective function. Time states the start time of a certain job serviced by a certain machine. For simplicity, we assume that time is discrete represented by an integer type. Therefore, release time, due time, processing time and start time will be considered as integer.

Before we present the formal specification of the underlying domain theory for job-shop scheduling problems, we first define the data type for these problems. These data types are used to deal with the properties of the scheduling problems. The primitive data types will be represented as

```
type Job-type = symbol
type Machine-type = symbol
type Time = integer
```

Other data types will be represented as a tuple based on the primitive data types such as

```
type Schedule-type =
        Tuple(Job : Job-type,
              Machine : Machine-type,
              Start-time : Time)
type Processing-Time-type =
        Tuple(Job : Job-type,
```

Machine : Machine-type,

Processing-time : Time)

type Release-Time-type =

Tuple(Job : Job-type,

Release-time : Time)

type Due-Time-type =

Tuple(Job : Job-type,

Due-time : Time)

type Processing-Order-type =

Tuple(Job : Job-type,

Machine : array of Machine-type)

Note that the maximum number of elements of the array is M, which is known. It is used to represent the order of the machine servicing a job.

The constraints now can be expressed as follows:

- Consistent-Release-Time(Schedule : set(Schedule-type))

  ≡ ∀sch ∈ Schedule then Release-Time-F(sch.Job) <= sch.start-time

- Consistent-Scheduling(Processing-Order : set(Processing-Order-types), Schedule : set(Schedule-type))

  ≡ (∀proc ∈ Processing-Order) ∀i ∃sch ∈ Scheduling ∋ proc.Job = sch.Job and proc.Machine[i] = sch.Machine

- Consistent-Machine-Assignment(Schedule : set(Schedule-type))

  ≡ (∀sch, sch' ∈ Schedule) (sch.Machine = sch'.Machine ∧ Start-Time-F(sch.Job) <= Start-Time-F(sch'.Job) <= Start-Time-F(sch.Job) + Processing-Time-F(sch.Job)) → sch = sch'

- Consistent-Job-Assignment(Schedule :set(Schedule-type))

≡ (∀sch, sch' ∈ Schedule) (sch.Job = sch'.Job ∧ Start-Time-F(sch.Job) <= Start-Time-F(sch'.Job) <= Start-Time-F(sch.Job) + Processing-Time-F(sch.Job)) → sch = sch'

- Consistent-Ordering(Schedule : set(Schedule-type))

≡ (proc ∈ Processing-time) ∧ (sch, sch' ∈ Schedule ∋ sch.Job = sch'.Job = proc.Job ∧ proc.Machine[i] = sch.Machine ∧ proc.Machine[j] = sch'.Machine) if i < j → Start-Time-F(sch.Job) < Start-Time-F(sch'.Job)

Deriving these constraints is straightforward. Furthermore, the objective can be an objective function mentioned previously.

Now we present the formal specification of a job-scheduling problem as follows:

```
function JSPS
        (Job : set(Job-type),
         Machine : set(Machine-type),
         Processing-Order : set(Processing-Order-type))
    returns {sch : set(Schedule-type) |
        Consistent-Release-Time(Schedule)
        Consistent-Scheduling(Processing-Order, Schedule)
        Consistent-Machine-Assignment(Schedule)
        Consistent-Job-Assignment(Schedule)
        Consistent-Ordering(Schedule)}
```

This specification specifies a function called JSSPS (Job-Shop Scheduling Problem Scheduler), that needs three inputs: a set of job, a set of available machines, and a set of all tasks to be performed and their order. This function returns a schedule consisting of a job j, a machine m, and a start time t of the job j serviced by the machine m. The schedule must satisfy all constraints given. The constraints are defined as above and transformed into, for example

function Consistent-Release-Time

(Schedule : set(Schedule-type)) : boolean

= ∀sch in Schedule ⇒

(Release-Time-F(sch.job) <= sch.start-time)

Here Release-Time-F is a function defined as follows:

function Release-Time-F

(job : Job-type) : integer

= (rt : Release-Time-type)

(rt.Job = job) ⇒ rt.time

This function takes input job of a job-type and returns a value of an integer type, which is the release time of the job.

Decomposing an Algorithm for Job-Shop Scheduling Problems Using KIDS

To compute a schedule for a given formal specification of job-shop scheduling problem, we can approach two different ways. They are local and global approaches. A local approach pays attention to individual schedules and finds a similarity relationship among them. It starts from an initial schedule and iteratively improves the schedule by considering the neighboring to complete the schedule. One major problem with a local approach is that it may fall into a local solution. Another approach is a global approach. A global approach pays attention to a set of schedules. An optimal schedule is search by iteratively splitting an initial set of schedules into subsets until the feasible schedule is easily extracted. Some examples of global approaches are backtracking, heuristic search, branch-and-bound.

In this research, we will apply a global approach to job-shop scheduling problems. As mentioned previously, KIDS has specialized design tactics for creating algorithms of various kinds. A

global approach is one of them. A global search algorithm generalizes the computational paradigms of binary search, backtracking, branch-and-bound, constraint satisfaction, heuristic search, and others. The basic idea is to represent and manipulate sets of candidate solutions. The principal operations are to extract candidate schedules from a set and to split a set into subsets. Derived operations include various filters which used to eliminate sets containing no feasible schedules. Global search algorithms start from an initial set containing all feasible schedules to a given problem instance and repeatedly extract schedules, split set into subsets, and eliminate via filters until no set remains to be split. The process is often described as a tree search in which a node represents a set of candidate schedules and an arc represents the split relationship between set and subset. The filters serve to prune off branches of the tree that cannot lead to feasible schedules. Those are to raise the level of efficiency of the program.

In solving job-shop scheduling problems using KIDS, a collection of candidate schedules is often infinite and they are very seldom represented extensionally even when finite. Therefore, the intuitive notion of global search can be formalized as the extensions of a problem theory with an abstract data type of an intensional representation called a space descriptor. In addition to the extraction and splitting operations mentioned above, the types also include a satisfaction predicate to determine when a candidate solution is in the set denoted by a descriptor. It is usually the term space or subspace used to denote both descriptor and the set that it denotes.

Formally, a gs-theory consists of the following structure [Smith, 1987], where D is the input domain, R is the output domain, I is the input condition, O is the output condition, Satisfies is the denotation of the descriptor, and Extract is the extractor of schedules from the spaces,

Sorts

$$D, R, \vec{R}$$

Operations

$I : D \rightarrow \text{Boolean}$

$O : D \times R \rightarrow \text{Boolean}$

$\vec{I} : D \times \vec{R} \rightarrow \text{Boolean}$

$\text{Satisfies} : R \times \vec{R} \rightarrow \text{Boolean}$

$\text{Split} : D \times R \times \vec{R} \rightarrow \text{Boolean}$

$\text{Extract} : R \times \vec{R} \rightarrow \text{Boolean}$

Axioms

GSO. $I(x) \Rightarrow \vec{I}(x, \vec{r}_0(x))$

GS1. $I(x) \wedge \vec{I}(x, \vec{r}) \wedge \text{Split}(x, \vec{r}, \vec{s}) \Rightarrow \vec{I}(x, \vec{s})$

GS2. $I(x) \wedge O(x, z) \Rightarrow \text{Satisfies}(z, \vec{r}_0(x))$

GS3. $I(x) \wedge \vec{I}(x, \vec{r}) \Rightarrow (\text{Satisfies}(z, \vec{r}) \Rightarrow \exists \vec{s} (\text{Split}^*(x, \vec{r}, \vec{s}) \wedge \text{Extract}(z, \vec{s})))$

where $\vec{R}$ is the type of space descriptors; $\vec{I}$ defines legal space descriptors; $\vec{r}$ and $\vec{s}$ vary over descriptors; $\vec{r}_0(x)$ is the descriptor of the initial set of the candidate solutions; $\text{Satisfies}(z, \vec{r})$ means z is the set denoted by descriptor $\vec{r}$ or z satisfies the constraint $\vec{r}$ represents; $\text{Split}(x, \vec{r}, \vec{s})$ means that $\vec{s}$ is a subspace of $\vec{r}$ with respect to input x; $\text{Extract}(z, \vec{r})$ means that z is directly extractable from $\vec{r}$. Axioms GSO asserts $\vec{r}_0(x)$ is a legal descriptor, GS1 asserts legal descriptors split into legal descriptors and Split induces a well founded ordering on spaces, GS2 provides the denotation of the initial descriptor, meaning all feasible solutions contained in the initial space, and GS3 provides the denotation of arbitrary descriptor $\vec{r}$: an output object z is in the set denoted by $\vec{r}$ if and only if z can be extracted after finitely many applications of Split to $\vec{r}$ where

$$\text{Split}^*(x, \vec{r}, \vec{s}) = \exists k \, (\text{Split}^k(x, \vec{r}, \vec{s}))$$

and

$$(\text{Split}^0(x, \vec{r}, \vec{t})) = (\vec{r} = \vec{t})$$

and for all k,

$$\text{Split}^{k+1}(x, \vec{r}, \vec{t}) = \exists \vec{s}\ (\text{Split}(x, \vec{r}, \vec{s}) \wedge \text{Split}^k(x, \vec{s}, \vec{t})).$$

Note that all variables are assumed to be universally quantified unless explicitly specified otherwise.

The initial schedule is just an empty schedule. This initial schedule is extended by first assigning a job into a machine at a certain start time. Given an assignable tuple (job, machine, start time), Split will attempt to extend a schedule with the tuple and check its legal descriptor. This process is repeatedly attempted until all jobs scheduled. Of course, it is possible that this assignment causes impossible to complete a feasible schedule. In tree search algorithm it is important to avoid such a violation as early as possible. Next we will examine this.

Beside those assignments, there are other derived operations that could lead to produce an efficient algorithm. These derived operations are called filters. In KIDS, filters play an important role to produce an efficient code. They are crucial to the efficiency of a global search algorithm and correspond to the notion of pruning branches in backtrack algorithm and pruning lower bounds and dominance relations in branch-and-bound. A feasibility filter

$$\psi : D \times R \rightarrow \text{Boolean} \quad \ldots\ldots (1)$$

is used to eliminate nodes from further processing. The ideal feasibility filters decide the question "Does there exist a feasible solution in space $\vec{r}$?" [Smith, 1990a] or to be more precise,

$$\exists(z \in R)\ (\text{Satisfies}(z, \vec{r}) \wedge O(x, z)). \quad \ldots\ldots (2)$$

Using (2) directly is too costly. Usually, other approaches are used such as:

- necessary feasibility filters $\Psi(x, \vec{r})$:

  $$\exists(z \in R)\ (\text{Satisfies}(z, \vec{r}) \wedge O(x, z)) \Rightarrow \Psi(x, \vec{r}).$$

  They only eliminate spaces not containing solutions, so they are generally useful but hard to find;

- sufficient feasibility filters $\psi(x, \vec{r})$:

$\psi(x, \vec{r}) \Rightarrow \exists(z \in R)$ (Satisfies(z, $\vec{r}$) $\wedge$ O(x, z)).

They are mainly used when only one solution is looked for; and

- heuristic feasibility filters bearing other relationships to $\exists(z \in R)$ (Satisfies(z, $\vec{r}$) $\wedge$ O(x, z)). These filters offer no guarantee, but a fast heuristic approximation to it may have the best performance in practice.

Given a global search theory G and a necessary filter ϕ, Smith proved that the following program specification F is consistent [Smith, 1987]:

function F(x : D) : set(R)

where I(x)

returns {z | O(x, z)}

= if $\phi(x, \vec{r}_0(x))$ then F_gs(x, $\vec{r}_0(x)$)

else {}

where

function F_gs(x : D, $\vec{r}$: $\vec{R}$) : set(R)

returns {z | Satisfies(z, $\vec{r}$) $\wedge$ O(x, z))

= {z | Extract(z, $\vec{r}$) $\wedge$ O(x, z)}

$\cup$ reduce($\cup$, {F_gs(x, $\vec{s}$) | (Split(x, $\vec{r}$, $\vec{s}$) $\wedge$ $\phi(x, \vec{s})$)})

This abstract global search program works as follows. Program F on input x calls F_gs with initial space $\vec{r}_0(x)$ if filter ϕ holds; otherwise, no feasible solutions. Program F_gs unifies two sets:

(1) all solutions that can be directly extracted from the space $\vec{r}$, and

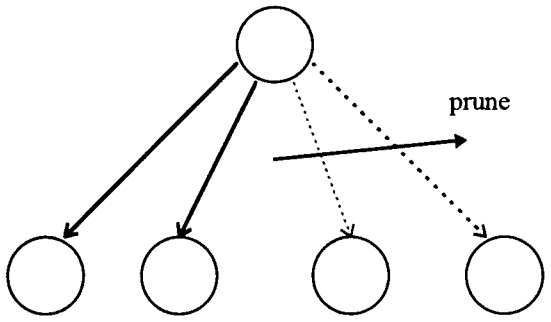
(2) the union of all solutions found recursively in spaces $\vec{r}$ that are obtained by splitting $\vec{r}$ and preserve the filter.

In other words, the global search unifies all solutions found at the current node with the solutions found at decedents. Note that filter ϕ is an input invariant in F_gs.

As mentioned above, it is important to detect a violation as early as possible during an extension of a partial schedule to lead an impossible feasible schedule. To detect such a violation, a necessary feasibility filter will be used. In this case we use a contrapositive of the necessary filter. That is, if $\neg\psi(x, \vec{r})$, then there are no feasibility schedule in $\vec{r}$. Therefore, it can be eliminated from the search space. So, the global search will apply this filter by testing $\psi(x, \vec{r})$ at any node that it visits and prune all nodes that $\psi(x, \vec{r})$ is false.

One way to get a pruning test is to initiate (2) with Satisfy and output condition O and use the inference system to derive a necessary condition.

Below will show a simple example of this pruning technique. Consider the following tree.



At the root, there are four successor nodes. It means there are four ways to assign a job to a machine. However, by further investigation of the necessary condition for the successor node, that there may be some assignments not possible. For example, the assignment is not possible due to the violation of the release time of the jobs. In this case, the whole subtree rooted at the successor node is pruning without further investigation. These kinds of tests can lead to early detection of infeasible schedules.

The above example involves the Consistent-Release-Time constraint. Another example is to examine constraint Consistent-Ordering. In extending a partial schedule, some nodes could be eliminated. The elimination involves the Consistent-Ordering constraint. Consider the following scenario. When a global search attempts to examine the successor of the current nodes, it is necessary to test whether the successor node will violate condition Processing-Ordering. If it is, that should be eliminated and there is no need to consider. In other words, we can eliminate all nodes that violate the constraint. Therefore, these kinds of tests could result an efficient code.

By early detection such pruning, the search space is converged faster. So, it is important to derive such necessary conditions.

Another way to get an efficient code is to perform constraint propagation. This technique is a more general way. It is essential for an early detection of infeasibility.

As mentioned before, a search space can be viewed as a tree. Each node is a data structure denoting a set of candidate solutions. A node is rooted at a node. It is called a parent. So, the parent node denotes the set of all candidate schedules found in the tree.

As contrast with pruning which removes a node from further consideration, constraint propagation changes the space descriptor so that it denotes a smaller set of candidate schedules. Its effect is to sharpen the subspace descriptor. In result, the descriptor becomes tight and exposing possibly infeasibility.

Constraint propagation is based on the concept of cutting constraints. Let $\vec{r}$ be a subspace descriptor in a global search tree denoting a subspace S = subspace($\vec{r}$) and let c(z, $\vec{t}$) be a cutting constraint where z is a candidate solution and $\vec{t}$ is an arbitrary subspace descriptor. Define an operation Cutting that takes $\vec{r}$ and c and generate a new descriptor $\vec{s}$ such that

$$\text{Cutting}(\vec{r}) = \vec{s} \Leftrightarrow \text{subspace}(\vec{s}) = \{z \mid z \in \text{subspace}(\vec{r}) \wedge c(z, \vec{r})\}$$

This operation will throw away space that does not contain the feasible solution. By applying this operation over and over, the space descriptor becomes smaller and smaller until no more space to throw away. This operation is similar to a fixpoint operator:

$$\text{Cutting}(\vec{s}) = \vec{s}$$

In a fixpoint operator, we repeatedly apply the operator until there is no more or a little changing in the operand.

This constraint propagation is used to optimize algorithm to find a tight lower bound on the cost of optimal solution for given objective functions [Nemhauser and Wosley, 1988].

To derive a cutting constraint is similar to deriving pruning. That is to find a candidate schedule that is feasible:

$$\forall(x \in D, \vec{r} \in \vec{R}, z \in R)\ (\text{Satisfies}(z, \vec{r}) \wedge O(x, z)) \Rightarrow \Psi(x, z, \vec{r}) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ (3)$$

It means that $\Psi(x, z, \vec{r})$ is a necessary condition to extend the candidate schedule for a given output condition.

Now we apply the constrapositive to this to determine the feasibility of the candidate schedule. That is, if $\neg\Psi(x, \vec{r})$ then z is not a feasible schedule that extends the previous schedule. Therefore, it can be pruned.

So, we may incorporate the derived necessary condition $\Psi(x, \vec{r})$ into $\vec{r}$ to obtain a new space descriptor to get an efficient code.

Concluding Remark

In this research we examine job-shop scheduling problems. After reviewing some literature, we come up with components of job-shop scheduling problems. Then we formulate a domain theory for the job-shop scheduling problems. This domain theory defines the basic concept of the problems in terms of functions, data types, and law, and inferences rules to reasoning about the concept.

We develop a formal specification for the domain theory. The formal specification serves to define the problem for which we desire an efficient computational solution in terms of the underlying domain theory. The problem is defined by means of functional constraint on the input/output behavior.

We identify some of heuristic filter that can be used to speed the process of finding the solutions. There are two main filters presented: pruning and constraint propagation. Pruning serves to eliminate all nodes that do not contain any feasible solution. Constraint propagation can be used to optimize algorithm to find a tight lower bound on the cost of optimal solution for a given objective function.

By using KIDS, we push in those two specification and design tactic to get high level programming. Later on, the high level programming is optimized and compiled into a correct and efficient executable code.

Due to an internal problem in installing KIDS in our system, we are not able to verify our result. However, we will continue to work in this problem until the system is properly working.

References

Bel, G., E. Bensana, D. Dubois, J. Erschler, and P. Esquirol. 1989. "A knowledge-based approach to industrial job-shop scheduling." In *Knowledge-Based System in Manufacturing*, Andrew Kusiak (ed.), Taylor & Francis, pp. 207-246, London.

Dorn, J. and R. Shams. 1991. "An expert system for scheduling in a steelmaking plant." In *Proceedings of the World Congress on Expert System*, pp.395-404, Pergamon Press, New York.

Fox, B. R., and K. G. Kempf. 1984. "ISIS – A knowledge-based system for factory scheduling." *Expert Systems*, Vol.1, No. 1, pp. 25-49.

Fox, B. and K. Kempf. 1985. "Opportunistic scheduling for robotic assembly." In *Proceedings 2nd IEEE Inter. Conf. Robotics and Automation*, pp. 880-889, IEEE Computer Society Press, Los Alamitos, CA.

Fox, M., N. Sadeh, and C. Baykan. 1985. "Constrained heuristic search." In *Proceedings of International Joint Conf. On Artificial Intelligence*, pp. 309-316, AAAI Press, Menlo Park, CA.

Fox, M. 1987. *Constrained-Directed Search: A Case Study of Job-Shop Scheduling*. Morgan Kaufmann, San Fransisco, CA.

Gary, K., R Uzsoy, S. P. Smith, and K. G. Kempf. 1992. "Accessing the quality of production schedules." In *Intelligent Scheduling Systems*, W. Scherer and D. Brown (eds.), Kluwer Academic Publishers, Norwell, MA.

K. G. Kempf. 1989. "Manufacturing scheduling – Intelligently combining existing methods." In *Proceedings AAAI, Stanford Spring Symposium*, pp. 51-55, AAAI Press, Menlo Park, CA.

K. G. Kempf. 1989. "Scheduling wafer fabrication – The intelligent way." *SME Electronics in Manufacturing*, Vol. 4, No.3, pp. 1-3.

Nemhauser, G. and L. Wosley. 1988. *Integer and Combinatorial Optimization*. John Wiley and Sons, Inc., New York, NY.

Numao, M. and S. Morishita. 1989. "A scheduling environment for steel-making process." In *Proceedings of the 5th Conference on Artificial Intelligence Applications*, pp. 279-286, IEEE Computer Society Press, Los Alamitos, CA.

Rinnoy Kan, A. 1976. *Machine Scheduling Problems*. Martinus Nijhoff, Hague, Netherlands.

Pinedo, M. 1995. *Scheduling: Theory, Algorithms, and Systems*. Prentice Hall, Englewood Cliffs, NJ.

Smith, D. 1987. *Structure and Design of Global search Algorithm.* Technical Report KES.U.87.12, Kestrel Institute, November, 1987.

Smith, D. 1990a. "KIDS: A semi-automatic program development system." In *IEEE Transaction on Software Engineering Special Issue on Formal Methods in Software Engineering*, Vol. 16, No. 9 (September), pp. 880-889.

Smith, D., and M. Lowry. 1990b. "Algorithm theories and design tactics." *Science of Computer Programming*, Vol. 14, No. 2-3, pp. 305-321.

Smith, D. 1991. "KIDS: A knowledge-based software development system." In *Automatic Software Design*, M. Lowry and R. McCartney (eds.), pp. 880-889, AAAI/MIT Press, MA.

Smith, D., and E. Parra. 1993. "Transformational approach to transportation scheduling." In *ARPA/RL Knowledge-Based Planning and Scheduling Initiative: Work-shop Proceedings*, February, pp. 205-216.

AN ANALYSIS OF THE ADAPTIVE DISPLACED PHASE CENTERED ANTENNA SPACE-TIME PROCESSING ALGORITHM

Rick S. Blum

Assistant Professor

Department of Electrical Engineering and Computer Science

Lehigh University

Bethlehem, PA 18015

Final Report for:

Summer Research Extension Program

Sponsored by:

Air Force Office of Scientific Research, Bolling AFB, Washington DC

and Rome Laboratory

January 1997

An Analysis of the Adaptive Displaced Phase Centered Antenna Space-Time Processing Algorithm

Rick S. Blum

Assistant Professor

Department of Electrical Engineering and Computer Science

Lehigh University

Abstract

Researchers have developed and examined Space-Time Adaptive Processing (STAP) schemes to cope with the clutter spectral spreading that occurs for a radar mounted on a moving platform. Analysis shows these schemes have great potential. Unfortunately, much of the previous evaluation of STAP algorithms was based on the assumption that accurate estimates of the interference-pulse-noise statistics are available which is usually unrealistic. In this report, performance evaluation is based on a highly non-homogeneous environment where interference-plus-noise statistics are unknown. Further, estimates which attempt to characterize the interference-plus-noise environment are obtained by probing nearby range cells which is typical in practice. Often, as we show, these estimates are very inaccurate. A general formulation of a STAP algorithm is defined and several specific cases are described and studied. Both simulated data and real measured radar data are used in the tests. The results indicate that STAP schemes can be developed which will perform well when operating with limited information and possibly mismatched estimates of the interference-plus-noise environment. Further development and study are needed to identify the best STAP schemes for this purpose.

An Analysis of the Adaptive Displaced Phase Centered Antenna Space-Time Processing Algorithm

Rick S. Blum

1 Introduction

In airborne radar, the detection of targets is often limited by ground clutter and other forms of interference. Platform motion causes Doppler shifts in the ground clutter that makes Doppler filtering alone ineffective. In such cases Space-Time Adaptive Processing (STAP) offers a potential solution.

STAP has been an active research topic for at least the last two decades. Much of the interest was generated by the results in [1] and [2]. Since then several algorithms have been proposed and evaluated using simulated radar data. With the recent improvements in phased array antenna and digital signal processing technology, a STAP-based radar system is becoming an attractive alternative for detection of small airborne targets in severe clutter, as compared to classical low-sidelobe beamforming [3].

Current STAP research efforts [4] are focused on, among other things, improved estimation of the clutter-plus-noise statistics, calibrated clutter measurements, real-time processing hardware development, and performance evaluation for the competing STAP approaches. The last one is the interest of this report. In most previous research, STAP schemes were evaluated using simulated data or by manipulating stationary platform measurements to simulate motion. While simulated data is very useful in the development and analysis of algorithms, a more complete evaluation includes using actual recorded radar data. Thus, in this report, we compare various STAP schemes using both simulated data and actual measured airborne data. A general STAP processing approach, which includes most linear processing schemes, is developed. This should be useful in designing robust STAP algorithms which is an important topic for future research.

In section 2 we describe the system under consideration and we provide an introduction to STAP. In Section 3 we define a general STAP scheme and give a detailed description of several specific approaches. In Section 4 we present comparison results which utilize simulated radar data. Comparison results based on measured airborne radar data are presented in Section 5. Conclusions are given in Section 6.

2 System Overview

A radar operates by transmitting energy into the environment and obtaining information concerning the location of objects by detecting reflections of the transmitted energy [5]. Detection of an object requires that the received energy from a reflection be larger than the background energy. There are many contributors to the background energy. The most fundamental one is the random energy fluctuations that result from the random motion of electrons. This contribution is often called noise. The presence of other reflectors, such as the ground, whose energy tends to obscure that of the reflector of interest, can also contribute to background energy. These contributors are denoted as clutter.

Thermal noise [6] is always present in electronic circuits. As the name implies, thermal noise is a function of temperature. For our purpose, an important aspect of thermal noise is that its power spectral density is constant over the bandwidth of typical radar receivers. The effective received noise power P_n due to the combined effects of the antenna and receiver is directly proportional to bandwidth. Thus

$$P_n = kT_sB \tag{1}$$

where k is Boltzmann's constant, B is the receiver bandwidth measured in Hz and T_s is the effective temperature of the receiving system.

In military radar applications, an adversary can decrease a radar's detection capability by inserting a signal in the bandwidth of the radar. This type of interference is known as jamming [3]. Typically, the energy received by the radar due to jamming is far greater than that of thermal noise, so the presence of the jammer can significantly decrease the

performance of the radar system. One significant difference between interference due to jamming and thermal noise is that interference due to jamming typically emanates from a single spatial angle, whereas thermal noise has essentially uniform density over spatial angles. Thus, modifying the antenna pattern to ensure a low response in the direction of the jamming will often decrease the effectiveness of the jamming [3].

Clutter [5] can be distributed in angle and range. Further, clutter can be distributed in Doppler frequency. For an airborne surveillance radar, the major source of clutter is ground clutter which is due to the backscattering of radiation from the ground. To detect a target in the presence of clutter with a stationary radar platform, a useful discriminant is that often targets have high velocity, and therefore high Doppler shifts, whereas ground clutter has zero or low velocity, and therefore zero or low Doppler shifts. Delay-line [7] cancelers can be easily configured to have nulls at zero Doppler to suppress this clutter. For moving platforms, Doppler shift is dependent on the aspect angle of a scatter relative to the radar look direction. As a result, this makes the Doppler spectral spread quite large. In order to cancel these aspect-dependent clutter returns, STAP has been found to be useful.

2.1 Radar System

The system under consideration is a pulsed Doppler radar residing on an airborne platform. As defined by the IEEE Standard Radar Definitions [5], a Doppler radar is one which utilizes the Doppler effect to determine the radial component of relative radar-target velocity or to select targets having particular radial velocities. It functions to enhance targets within a particular velocity band while rejecting clutter and other echoes outside the velocity band of interest. When a Doppler radar uses pulsed transmissions, it is called a pulsed Doppler radar.

We assume the radar transmits a coherent burst of M pulses at a constant pulse repetition frequency $f_r = 1/T_r$, where T_r is the pulse-repetition-interval (PRI). The time interval over which the waveform returns are collected is commonly referred as coherent-processing-interval (CPI).

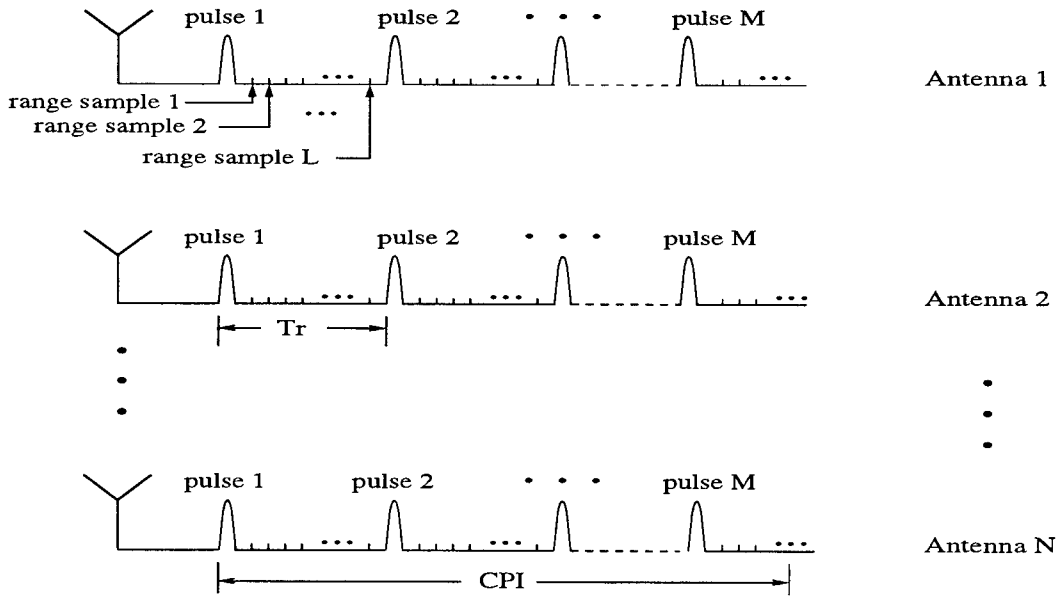


Figure 1: Structure of the observed signal returns.

In our analysis, the radar antenna used is a uniformly spaced linear array antenna with N identical elements (For the measured data cases we consider a two dimensional array.). These elements may be the beamformed columns of a rectangular planar array. It is also assumed that the radar array has a fixed transmit pattern. Each element of the array has its own down-converter, matched filter receiver, and A/D converter. For each PRI, L time samples are collected to cover the range intervals as illustrated in Fig. 1.

With M pulses and N antennas, the received data for one CPI comprises LMN complex baseband samples. This multidimensional data set, collected as shown in Fig. 1, is often referred to as a datacube. Denote the observation corresponding to the i^{th} antenna element at the j^{th} pulse for the k^{th} range cell as $x_{i,j,k}$. It is convenient to denote the part of the datacube which represents the k^{th} range cell of the datacube as

$$\mathbf{X}_k = [x_{1,1,k}, x_{2,1,k}, \ldots, x_{N,1,k}, x_{1,2,k}, \ldots x_{N,M,k}]^T \tag{2}$$

where a^T denotes the transpose of the vector a. We will refer to $\mathbf{X}_k$ as a space-time snapshot.

As in most analysis of radar systems [8], we assume we can decompose the samples

in the datacube as

$$\mathbf{X}_k = \alpha V(S) + X(C) \tag{3}$$

where $V(S)$ is the normalized target response, given as

$$V(S) = b(\varpi) \otimes a(\vartheta). \tag{4}$$

In (4), $\otimes$ denotes the Kronecker product,

$$b(\varpi) = [1, e^{j2\pi\varpi}, \ldots, e^{j(M-1)2\pi\varpi}] \tag{5}$$

is an $M \times 1$ temporal steering vector in which ϖ is the normalized target Doppler frequency as defined in [8], and

$$a(\vartheta) = [1, e^{j2\pi\vartheta}, \ldots, e^{j(N-1)2\pi\vartheta}] \tag{6}$$

is an $N \times 1$ spatial steering vector in which ϑ is the target spatial frequency as defined in [8]. The symbol $V(S)$ used in (4) is often called a target steering vector. In (3), α is an unknown constant and $X(C)$ denotes the additive interference-plus-noise returns. $X(C)$ usually consists of additive contributions of clutter, jamming and thermal noise. A typically model for each of the components of $X(C)$ is given in [8].

2.2 Space-Time Adaptive Processing

Most of the STAP schemes that have been suggested can be represented as an inner product of the conjugate of a weight vector w and the vector $\mathbf{X}_k$ which represents the snapshot of interest. This inner product

$$z = w^H \mathbf{X}_k \tag{7}$$

produces the complex quantity z whose magnitude is often compared to a threshold to make a decision. The weight vector w may depend on the estimated interference-plus-noise environment and on the target of interest.

One way to view a space-time processor is as a two-dimensional filter. Conventional schemes only use Doppler frequency selectivity and could produce a filtering action as shown

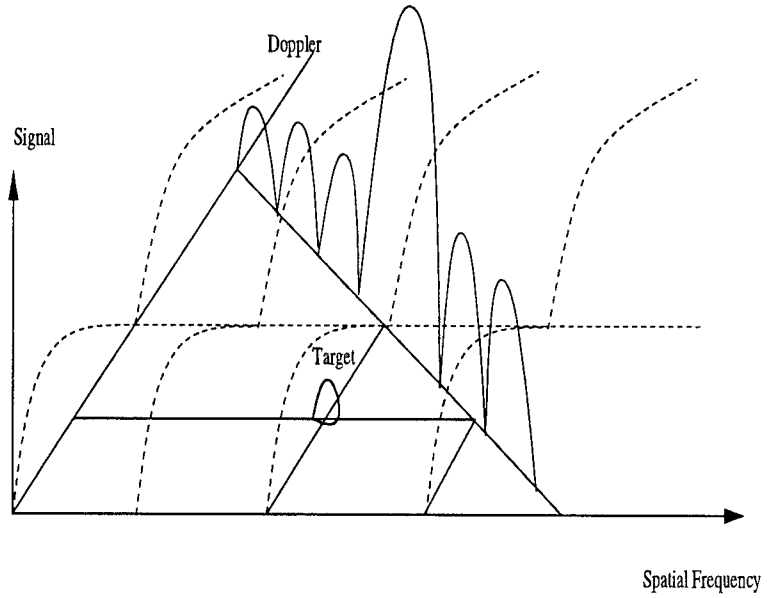


Figure 2: One-dimensional Doppler filter applied to ground clutter (adopted from [9]).

in Fig. 2. Fig. 2 shows the clutter ridge that is characteristic of ground clutter in airborne radar [8]. It is clear that although such a filter can cancel the clutter return due to the mainlobe (assumed to be at zero Doppler) of the antenna response, it is unable to cancel the clutter returns due to sidelobes. STAP schemes combine both the spatial and temporal information and are able to rotate the filter to produce a null along the clutter ridge as shown in Fig. 3. Ideally, the space-time processor provides coherent gain for a target while forming angle and Doppler response nulls to suppress interference. As the interference scenario is not known in advance, the weight vector must be determined in a data-adaptive way from the radar returns.

In the well-know sample matrix inversion (SMI) algorithm [2], which is a fully adaptive algorithm, the weight vector is given, to within a scale factor, by

$$w = \hat{R}^{-1} V(S) \tag{8}$$

where $\hat{R}$ is the estimated interference-plus-noise covariance matrix. The estimate is based on a set of reference data, typically chosen from the surrounding range cells. $V(S)$ is the normalized target response defined in (4).

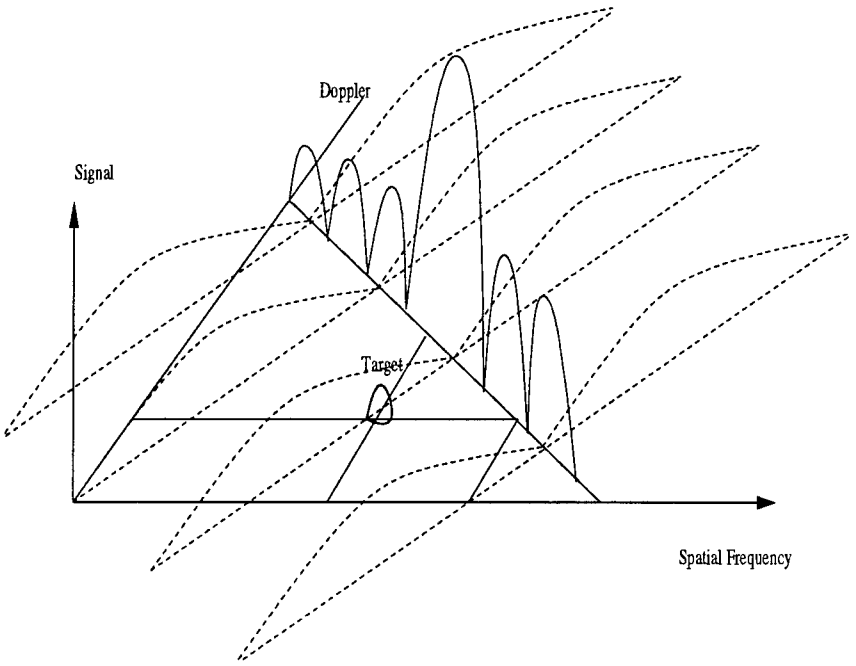


Figure 3: Space-time filter applied to ground clutter (adopted from [9]).

In the case where the interference statistics are known or the estimated covariance matrix is exactly equal to the true covariance matrix, SMI can achieve optimal performance. A fully adaptive STAP scheme is one that requires the formation of an NM by NM covariance matrix which can be a problem. Even for moderate M and N, the computational cost of the estimation and computation of $\hat{R}^{-1}$ becomes excessive in real-time implementations. As a result, reduced complexity approaches have been developed whose computational cost is substantially smaller. Some examples of reduced complexity approaches are given in the next section.

The schemes which estimate the interference-plus-noise statistics typically require a large set of independent and identically distributed (iid) reference data vectors to achieve an accurate estimation. This requirement may be unrealistic, since measurements [10] indicate that multichannel airborne radar clutter data is often severely non-homogeneous. For this reason the reference data set available for estimation of clutter statistics is usually quite small. Therefore it is important to know how different STAP algorithms perform for such cases.

A particular STAP algorithm, the adaptive displaced phase-centered antenna (ADPCA) algorithm, appears to provide benefits in some non-homogeneous environment cases where the interference statistics estimates may be inaccurate.

3 Some Reduced Complexity STAP Schemes

STAP is an active research area and new schemes are continually being developed. In order to compare schemes, a standard terminology is useful. Here, we will mainly follow the terminology used in [8]. We caution the reader that other terminology also appears in the literature. We first define a general formulation of a for STAP processing approach which encompasses most of existing STAP schemes. We limit consideration to those schemes which linearly combine the space-time observations. Next we describe six specific approaches which are

- Adaptive displaced phase-centered antenna (ADPCA)
- Factored post-Doppler (post-Doppler adaptive beamforming)
- Element-space pre-Doppler
- Beam-space pre-Doppler
- Beam-space post-Doppler
- Joint-domain localized (JDL) approach

each of which are included in the general formulation.

3.1 General STAP Approach

Consider the transformations

$$\widetilde{X}_k(p) = (A_p \otimes B_p)^H \mathbf{X}_k; \quad p = 0, 1, 2, \ldots, P-1 \tag{9}$$

where $\mathbf{X}_k$ is the space-time snapshot from the kth range cell and A_p and B_p are scheme-dependent matrices. The operations in (9) can be interpreted as a pre-processor applied to

the received signals. This pre-processing generates data for the adaptive processing to follow. Note that P vectors are produced by the operations in (9). Typically, the pre-processing in (9) performs a coordinate transformation and a selection operation.

We describe the adaptive processing on the pth vector produced by (9) as

$$\tilde{y}_k(p) = S^H \tilde{R}_k^{-1}(p) \widetilde{X}_k(p) \tag{10}$$

where

$$\tilde{R}_k(p) = \frac{1}{Q} \sum_{i=k-Q/2, i\neq k}^{k+Q/2} \widetilde{X}_i(p) \widetilde{X}_i(p)^H \tag{11}$$

and S is a scheme-dependent steering vector. $\tilde{R}_k(p)$ is the interference-plus-noise covariance matrix estimated from Q adjacent range cells. Note that (10) resembles the SMI scheme defined in (7) and (8). Further, based on accepted principles, the covariance matrix estimation of an $r \times r$ matrix like $\tilde{R}_k(p)$ nominally requires $Q = 2r$ iid secondary data.

In different schemes, $\tilde{y}_k(p)$ may or may not be the final output of interest. If $\tilde{y}_k(p)$ is the final output of interest, its magnitude will be compared to a threshold to decide if signal is present. For cases where $\tilde{y}_k(p)$ will be processed further, we assemble the complex outputs from each adaptive processor into a $P \times 1$ vector as

$$\tilde{Y}_k = [\tilde{y}_k(0), \tilde{y}_k(1), \cdots, \tilde{y}_k(P-1)]^T \tag{12}$$

and compute

$$\tilde{z}_{k,m} = f_m^H \tilde{Y}_k \tag{13}$$

which we call post-processing (after adaptive processing). Typically, f_m is the mth column of a $P \times P$ filter matrix F, and $\tilde{z}_{k,m}$ is the final output whose magnitude will be compared to a threshold to produce a decision. A diagram of the complete processing flow is shown in Fig. 4.

3.2 Adaptive Displaced Phase-Centered Antenna

ADPCA is a low complexity alternative to fully adaptive schemes like SMI. ADPCA uses adaptive processing with K_t (typically 2 or 3) pulses at a time rather than all the pulses of

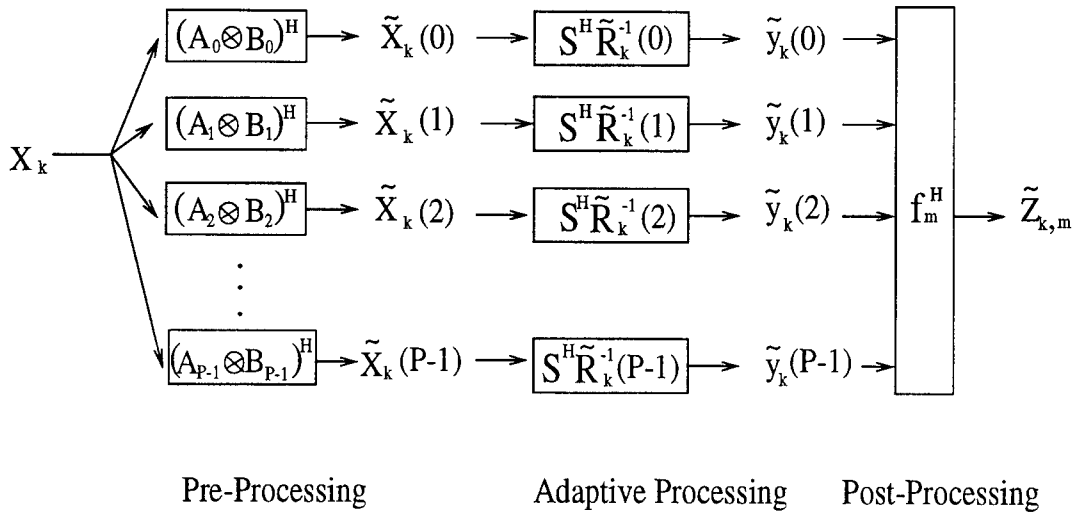


Figure 4: Processing flow for a general STAP scheme.

the CPI. To be more precise, define a set of P sub-CPIs $\widetilde{X_k}(p), p = 0 \cdots P-1$ in the kth snapshot. Each sub-CPI contains possible signal returns from K_t pulses and all N elements.

Fig. 5 shows two different ways to form the sub-CPIs. As indicated in Fig. 5, implementation (a) does not overlap pulses. Given M pulses in a CPI where M can be divided by K_t, implementation (a) generates $P = M/K_t$ sub-CPIs. The 0th sub-CPI consists of pulses $0, \ldots, K_t - 1$ and the pth sub-CPI consists of pulses $pK_t, \ldots, pK_t + K_t - 1$. Implementation (b) forms the sub-CPIs by using the same pulse returns in several sub-CPIs. Given M pulses in a CPI, implementation (b) generates $P = M - K_t + 1$ sub-CPIs. The 0th sub-CPI consists of pulses $0, \ldots, K_t - 1$ and the pth sub-CPI consists of pulses $p, \ldots, p + K_t - 1$. In Fig. 5, K_t is set to 3 and in implementation (b), neighboring sub-CPIs overlap 2 pulses. Of course, other overlaps are possible.

The pre-processing we have just described can be put into the framework of (9). B_p is set to I_N which is an $N \times N$ identity matrix and

$$A_p = \begin{bmatrix} 0_{p(K_t-h)\times K_t} \\ I_{K_t} \\ 0_{(M-K_t-pK_t+ph)\times K_t} \end{bmatrix} \tag{14}$$

Where the notation $0_{l\times m}$ refers to an $l \times m$ matrix of zeros. h indicates the number of pulses

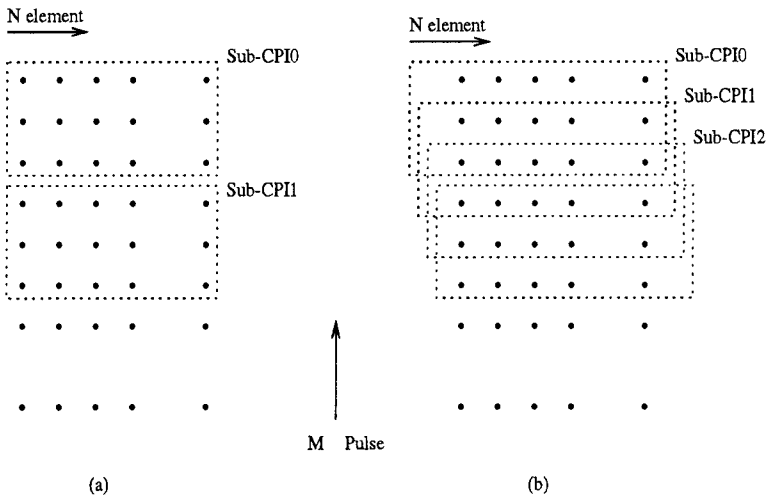


Figure 5: ADPCA sub-CPI formation.

which are overlapped. In implementation (a) h is set to be zero and in implementation (b) h is set to be $K_t - 1$. A_p is an $M \times K_t$ selection matrix.

The adaptive processing in ADPCA is described by (10) with the steering vector

$$S = S_t \otimes S_s \tag{15}$$

where S_s is the $N \times 1$ spatial steering vector as in (6), S_t is a $K_t \times 1$ vector, which is composed of the binomial coefficients, with each coefficient altered in sign (start with positive). As a particular example, we have

$$S_t = (1, -2, 1)^T \tag{16}$$

for a three pulse case. If $\tilde{R}_k(p)$ is an identity matrix, and we consider steering to broadside, then application of the ADPCA steering vector has a simple interpretation. At each element, subtract the amplitude of neighboring pulses. Next subtract neighboring results and repeat this process until only a single output is obtained. Finally the outputs from each antenna are summed. It is clear that in this case ADPCA is implementing a pulse differencing scheme which will tend to "whiten" clutter present in the observations.

Typically, post-processing as described in (13) is employed in ADPCA. In ADPCA F is a matrix corresponding to a Doppler filter bank, and f_m is the mth Doppler filter. Typically F is DFT matrix and this Doppler processing can then be efficiently implemented

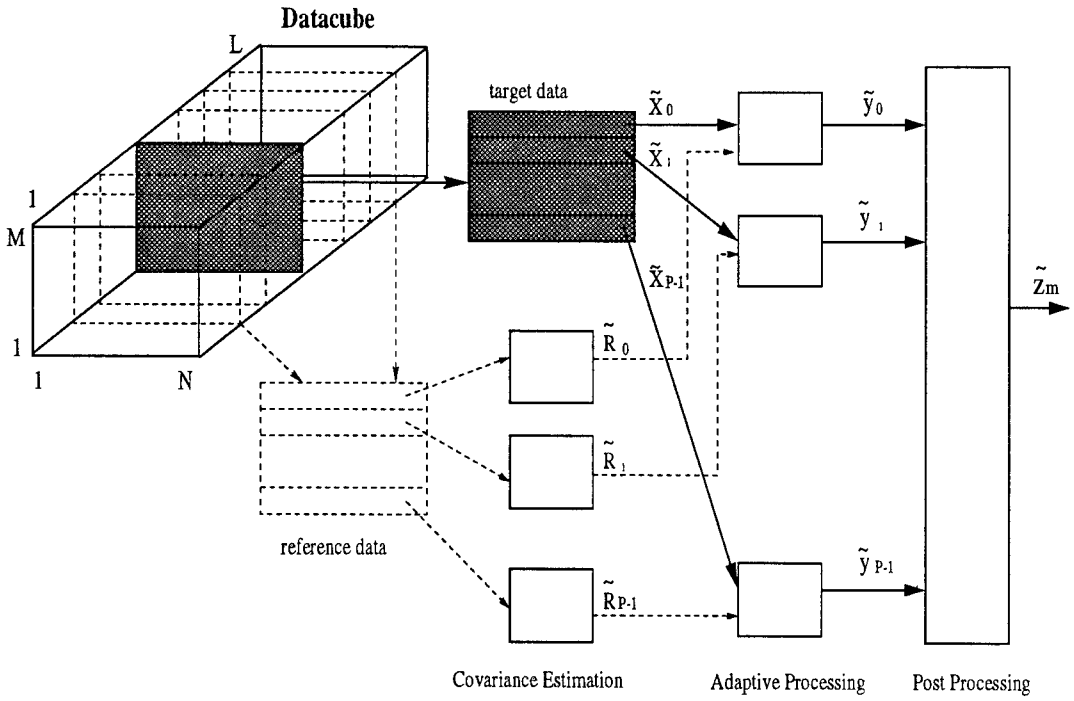


Figure 6: Block diagram illustrating the ADPCA algorithm.

by computing an FFT. If adaptive processing completely "whitens" the data, then FFT post processing is optimum. $|\tilde{z}_{k,m}|$, the final output for Doppler bin m, is compared to a threshold to make a decision if a target is present at this Doppler frequency.

Fig. 6 illustrates the principle of the ADPCA processor. One advantage of this approach is that it allows an accurate estimate of $\tilde{R}_k(p)$ to be obtained with only a small number of reference samples, since the dimensions of $\tilde{R}_k(p)$ are kept small. This can be quite important in practice due to the difficulty in obtaining a large set of homogeneous reference samples [10]. Further, if we assume stationarity returns over the CPI, we can use only one $\tilde{R}_k(p)$ for all p.

3.3 Factored Post-Doppler

In factored post-Doppler STAP [8], Doppler processing is first performed on each spatial channel resulting in a transformed signal matrix. Let the Doppler filter on each element be represented by f_p and for convenience collect the Doppler filters in the $M \times M$ matrix

$F_M = [f_0, f_1, \ldots, f_{M-1}]$. Then the pre-processing is described by (9) with $A_p = f_p$, $P = M$, and $B_p = I_N$. This pre-processing transforms the signal into Doppler space. In this case p indicates the index of Doppler bin in question. Next, the adaptive processing in (10) is employed with the steering vector S defined as in (6). As for most of the STAP schemes we discuss, tapering could be applied to the steering vector [8]. Post-processing is not usually employed, so $|\tilde{y}_k(p)|$ is compared to a threshold to test for a target in the pth Doppler bin.

The extended factored approach (EFA) [11] is a slight extension of the factored post-Doppler approach. In EFA, adaptive processing is applied to several adjacent Doppler bins instead of just one. Thus, the pre-processing performs both transformation and selection. In a case where the scheme adapts over 3 adjacent bins, the pre-processing can be described as in (9) with $A_p = J_p = [f_{p-1}, f_p, f_{p+1}]$. The other quantities are set the same as in factored post-Doppler STAP. Using EFA as opposed to post-Doppler STAP, will necessarily increase the size of the covariance matrices to be estimated and makes this approach closer to fully adaptive schemes like SMI.

3.4 Element-Space Pre-Doppler

In element-space pre-Doppler STAP [8], the adaptive processing considers only a few pulses at a time. Utilizing more than one pulse provides the temporal adaptivity required for clutter cancellation, while retaining full spatial adaptivity provides a means to handle jamming simultaneously. Clearly this approach is similar to ADPCA in structure. We begin with defining P sub-CPIs each containing signal returns from K_t successive pulses and all elements. As in ADPCA, one could utilize either an overlapped pulse configuration or a non-overlapped pulse configuration (as illustrated in Fig. 5). Thus, the pre-processing is described by (9) with A_p as described in (14) and with $B_p = I_N$. The adaptive processing is described by (10) with the steering vector S being a K_t-pulse, N-element normalized target response as in (4). Post-processing is usually employed to transform the output into Doppler space. In the standard approach, this post-processing is similar to what was described for ADPCA.

3.5 Beam-Space Pre-Doppler STAP

In beam-space pre-Doppler STAP [8], the dimensionality is reduced in two ways. First, the element data is pre-processed with an $N \times K_s$ beamformer matrix G to produce K_s beam outputs (see [8] for examples and further discussion of the choice of G). Second, only the beam outputs from a K_t-pulse sub-CPI are adaptively processed at one time. Typically $K_t \ll M$ and $K_s \ll N$ so that a significant reduction in problem size is achieved. This pre-processing is described by (9) with A_p as defined in (14), and $B_p = G$.

The adaptive processing is described by (10) with the beam-space steering vector S defined as

$$S = (I_{K_t} \otimes G)^H S_e \tag{17}$$

In (17), S_e is the steering vector for element-space pre-Doppler STAP. In the standard approach, the post-processing is the same as that described for element-space pre-Doppler.

3.6 Beam-Space Post-Doppler STAP

In the beam-space post-Doppler approach [8], the pre-processing is described by (9) with $A_p = J_p$ and $B_p = G$, where J_p is an $M \times K_t$ matrix of Doppler filters. A good example of J_p was that given when we introduced EFA. G is an $M \times K_s$ beamforming matrix similar to that described in Section 3.5. As in factored post-Doppler, the pre-processor transforms the data into Doppler space. The adaptive processing is as defined in (10) with the steering vector

$$S = (J_p \otimes G)^H V \tag{18}$$

where V is the M-pulse, N-element normalized target response as in (4). $\tilde{y}_k(p)$ in (10) is the final output for pth Doppler bin. Post-processing is not usually employed.

3.7 Joint-Domain Localized Approach

In the joint-domain localized (JDL) [12] approach, pre-processing transforms data from the space-time domain into the angle-Doppler domain. Then only a few angle-bins covering

angles near the target of interest are considered in the adaptive processing. Further, only a few Doppler bins adjacent to the Doppler bin of interest are adaptively processed. Thus, the pre-processor performs two-dimensional transformation and selection. Most conveniently, the transform is the two-dimensional DFT, and the selection picks out a local processing region (LPR) of width L_n in angle and L_m in Doppler.

More precisely, we can define this pre-processing using (9) with $A_p = [f_{M,1}, f_{M,2}, \cdots, f_{M,L_m}]$, where $f_{M,i}$ $i = 1 \ldots L_m$, are the columns of an $M \times M$ DFT matrix corresponding to the L_m Doppler bins in the LPR, and with $B_p = [f_{N,1}, f_{N,2}, \cdots, f_{N,L_n}]$, where $f_{N,i}$ $i = 1 \ldots L_n$, are the columns of an $N \times N$ DFT matrix corresponding to the L_n angle bins in the LPR. The adaptive processing can be described as in (10). For a uniform PRI and array spacing, the steering vector used in (10) has all its entries equal to zero except for the one corresponding to the angle and Doppler bin of the target. No post-processing is employed for JDL.

4 Performance Comparison Using Simulations

The performance comparison will be separated into two sections. In this section, theoretically based clutter and noise models are used to described the covariance matrix and hence generate data by computer simulation. A non-homogeneous environment is simulated by choosing an intentional statistics mismatch between the reference samples and the the samples taken from the cell-under-test. In the next section, actual measured radar data is used [10].

4.1 A Simple Model

Consider a simple model [12], which assumes that ground clutter is dominant over other sources of interference. Noise-plus-clutter observations are assumed to consist of additive contributions of noise and clutter and the noise and clutter are assumed to be statistically independent. Furthermore, the noise contribution to the noise-plus-clutter is assumed to

be Gaussian distributed and the noise observations at different antenna elements and in different pulses are assumed to be statistically independent.

The clutter contributions have a two dimensional power spectral density (psd) as described by [12]

$$P_c(f_t, f_s) = \sum_{d=1}^{K} \frac{\sigma_{c,d}^2}{2\pi\sigma_{ft,d}\sigma_{fs,d}} \exp\left[-\left(\frac{(f_t - f_{ct,d})^2}{2\sigma_{ft,d}^2} + \frac{(f_s - f_{cs,d})^2}{2\sigma_{fs,d}^2}\right)\right] \tag{19}$$

which is a function of Doppler frequency f_t and spatial frequency f_s. The psd in (19) consists of K Gaussian-shaped humps, the dth of which is centered at $(f_t, f_s) = (f_{ct,d}, f_{cs,d})$ and has amplitude $\sigma_{c,d}^2$ and a spread in angle and Doppler controlled by $\sigma_{ft,d}^2$ and $\sigma_{fs,d}^2$. The parameters in (19) are taken to model the clutter ridge observed in airborne radar. Using this model, we can easily generate mismatched reference data by manipulating parameters in (19). In our tests we frequently add one extra hump to the psd corresponding to either the cell-under-test or the reference data. For convenience of reference, we refer to this model as the simple model. Various experiments showed that how the mismatch is generated (whether the extra hump is added in the reference data or added in the cell-under-test and the exact location where the extra hump is added) will influence the relative performance of STAP schemes. Next we present some typical examples.

4.1.1 Case 1

Consider a case where the reference data can be described by the simple model with the parameters set as shown in Table 1 and with $K = 5$. The noise power in the sample observed at each antenna element and due to each pulse is taken to be 0.001 which gives the clutter-to-noise-ratio $CNR = 50\ dB$. We show the psd of clutter contributions in Fig. 7, where the highest peak is from the mainlobe and the rest of peaks represent sidelobes. This simplified clutter spectrum is useful for performance evaluations. It gives a simple representation of the clutter ridge. Assume that the cell-under-test has the same statistics except that its clutter psd includes an additional sixth hump with $\sigma_{c,6} = 2.0$, $\sigma_{ft,6} = 0.01$ and $\sigma_{fs,6} = 0.01$. We present three comparison results and in each of them the sixth hump is added at a

d	$\sigma_{c,d}$	$\sigma_{ft,d}$	$\sigma_{fs,d}$	$f_{ct,d}$	$f_{cs,d}$
1	0.5588	0.01	0.01	-0.35	-0.35
2	0.5588	0.01	0.01	-0.2	-0.2
3	9.9837	0.01	0.01	0.0	0.0
4	0.5588	0.01	0.01	0.2	0.2
5	0.5588	0.01	0.01	0.35	0.35

Table 1: Parameters of assumed psd for training samples.

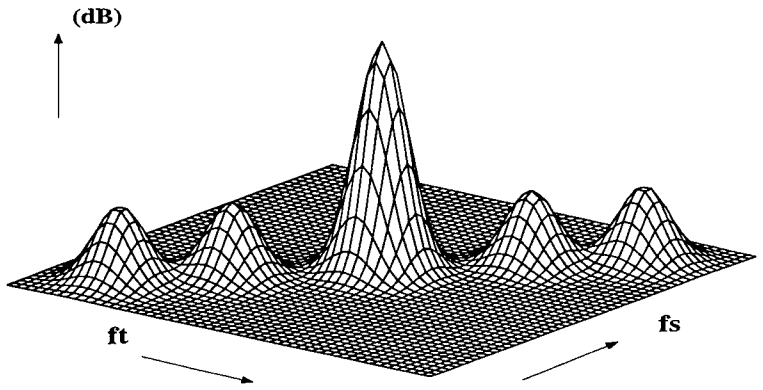


Figure 7: Clutter power spectral density for simple model.

different Doppler and spatial frequency. The sixth hump is placed along the clutter ridge at $f_{ct,6} = 0.1$ and $f_{cs,6} = 0.1$, $f_{ct,6} = 0.18$ and $f_{cs,6} = 0.18$, or $f_{ct,6} = 0.3$ and $f_{cs,6} = 0.3$. This type of mismatch could model a case where there is a large clutter return from a few discrete scatters which are not present in the reference data. Such cases have been observed in recently measured airborne data [14]. Alternatively the difference between the psd of the cell-under-test and the psd of the reference data could be the result of false target jamming [4]. For simplicity, a single target is assumed at $\vartheta = 0$ and $\varpi = 0.2$, where ϑ and ϖ are defined in (6) and (5).

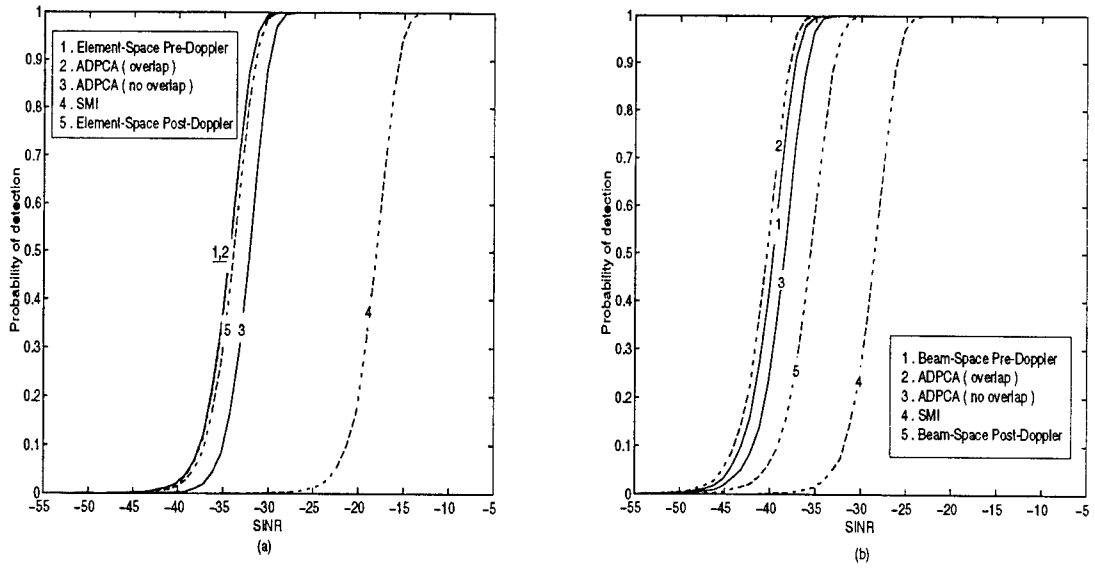


Figure 8: Performance comparison for simple model in case 1 with $f_{ct,6} = 0.1$ and $f_{cs,6} = 0.1$.

We compare the probability of detection of different STAP schemes as a function of signal-to-interference-plus-noise-ratio (SINR), as discussed in [15]. For each example, two groups of tests are conducted. In group (a), SMI, factored post-Doppler, element-space pre-Doppler, and ADPCA (with two different pulse grouping configurations as shown in Fig. 5), are compared for the case where the datacube consists of 2 elements and 12 pulses. In group (b), SMI, beam-space pre-Doppler, beam-space post-Doppler, and ADPCA (with two different pulse grouping configurations) are compared for the case where the datacube consists of 4 elements and 12 pulses. See Appendix A for the beamforming matrix, as well as for a summary of all the particular parameters chosen.

Fig. 8 through Fig. 10 illustrate the results. All the results were obtained by setting the true false alarm probability to be $P_f = 0.0001$, and using a Monte Carlo simulation with 10000 runs. Note that in group (b), the datacube includes more pulses than in group (a), so that detection performance is generally improved for all schemes.

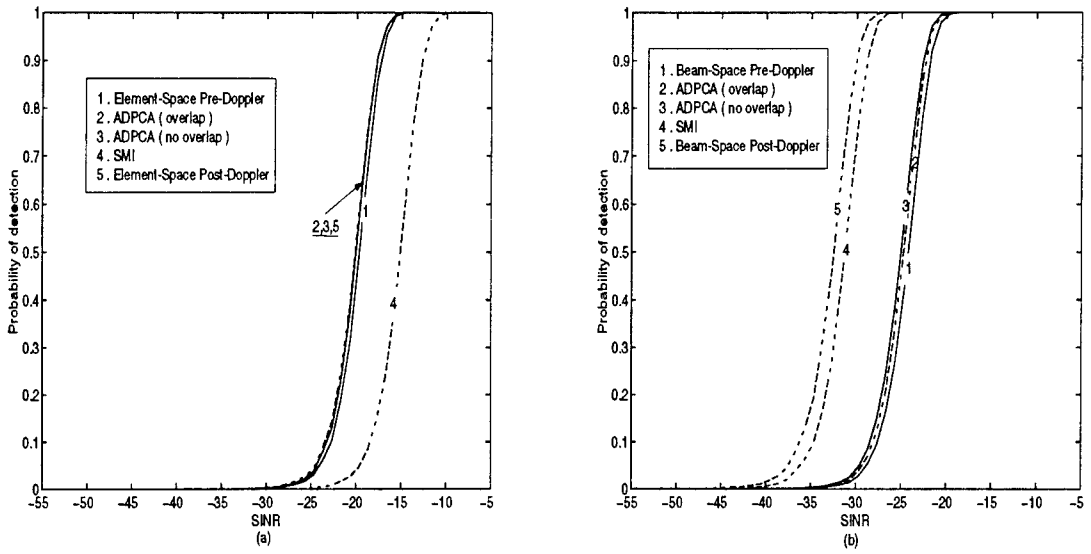


Figure 9: Performance comparison for simple model in case 1 with $f_{ct,6} = 0.18$ and $f_{cs,6} = 0.18$.

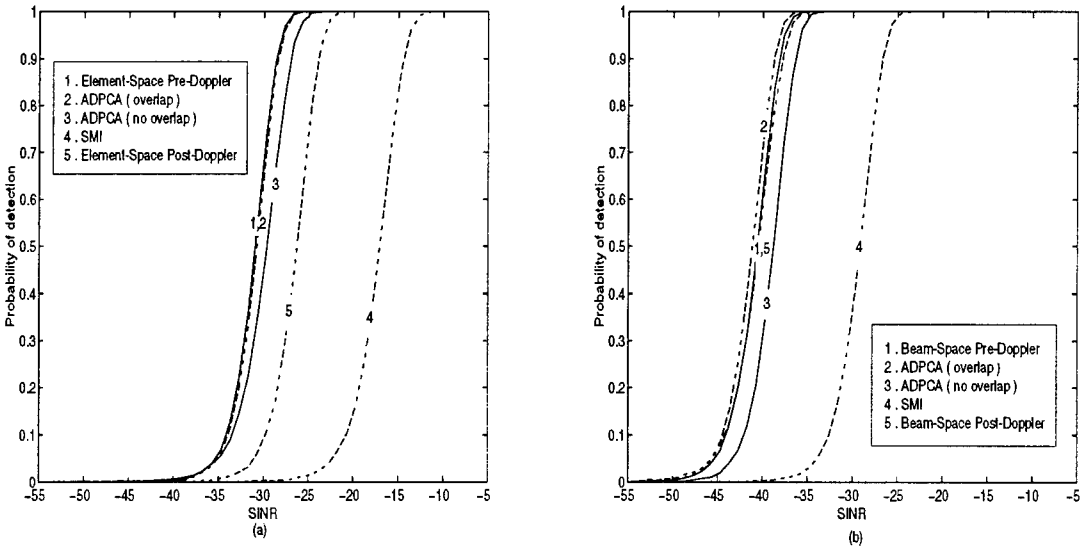


Figure 10: Performance comparison for simple model in case 1 with $f_{ct,6} = 0.3$ and $f_{cs,6} = 0.3$.

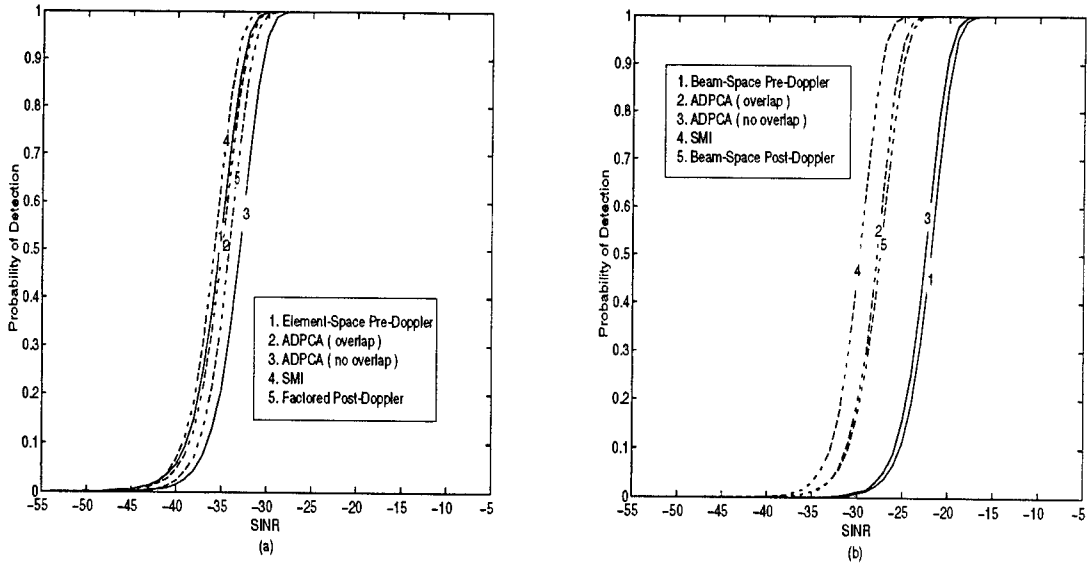


Figure 11: Performance comparison for simple model in case 2 with $f_{ct,6} = 0.1$ and $f_{cs,6} = 0.1$.

4.1.2 Case 2

Alternatively, mismatch may occur when the psd of the reference data contains an additional hump which does not exist in the psd of the cell-under-test. Assume that the psd for the clutter from the cell-under-test has the parameters in Table 1 and that the psd for the reference data contains an additional sixth hump with $\sigma_{c,6} = 4.0$, $\sigma_{ft,6} = 0.01$ and $\sigma_{fs,6} = 0.01$. Results are provided for examples where the extra hump in the psd of the reference data is located at $f_{ct,6} = 0.1$ and $f_{cs,6} = 0.1$, $f_{ct,6} = 0.18$ and $f_{cs,6} = 0.18$, and $f_{ct,6} = 0.3$ and $f_{cs,6} = 0.3$. As in case 1, we assume a single target at $\vartheta = 0$ and $\varpi = 0.2$. Fig. 11 through Fig. 13 illustrate the results. We conduct two groups of tests under the same condition as described in case 1.

4.2 SSC Model

The other model used, which we call the SSC model, is based on a more detailed description of the physical situation as described in [13]. In the SSC model, moving targets, jamming, receiver noise, and ground clutter are considered. The ground clutter model follows the

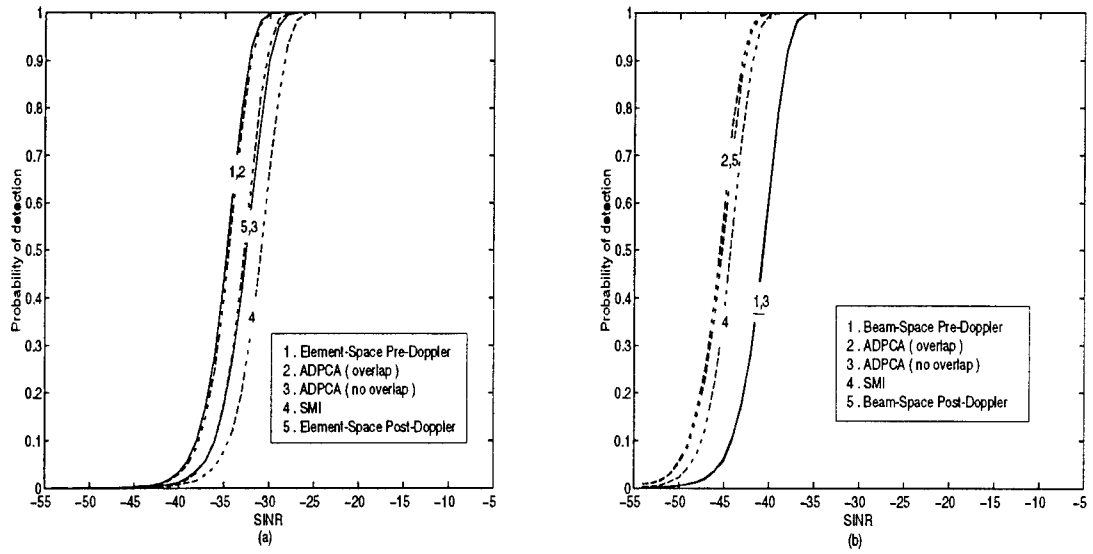


Figure 12: Performance comparison for simple model in case 2 with $f_{ct,6} = 0.18$ and $f_{cs,6} = 0.18$.

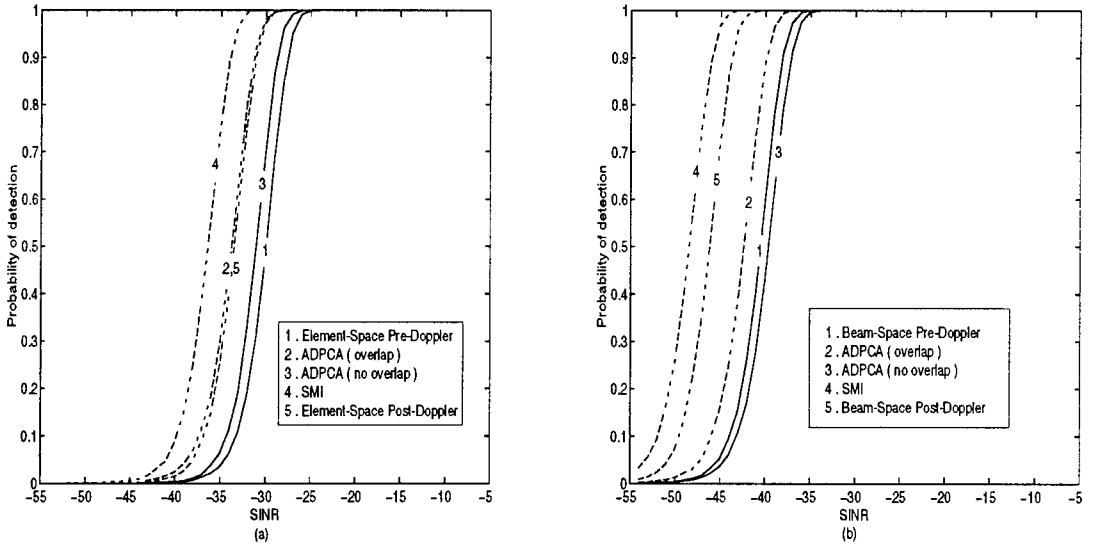


Figure 13: Performance comparison for simple model in case 2 with $f_{ct,6} = 0.3$ and $f_{cs,6} = 0.3$.

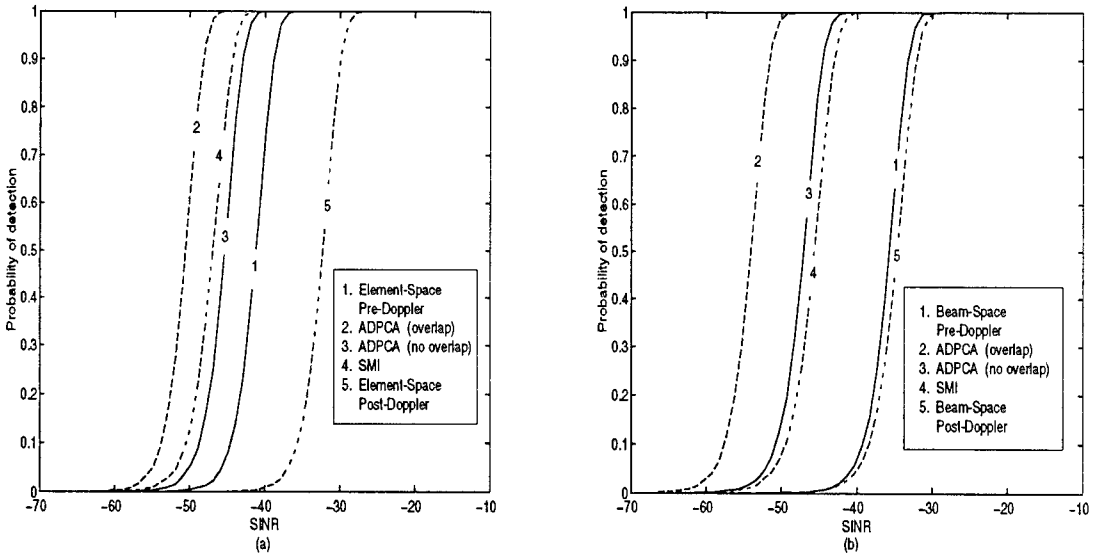


Figure 14: Performance comparison for SSC model with additional jamming in the psd of reference data

approach used in [8], where the contributions of many point scatters are added. By varying parameters in the SSC program, it is possible to model cases where there is a statistical mismatch between the cell-under-test and the reference data used to estimate $\hat{R}$.

For cases with mismatch between the statistics of the estimated and true ground clutter, we found results which were very similar to those found for the simple model. In contrast, here we show one case where the mismatch is caused by jamming. In this case, the psd of the data from the cell-under-test consists of only noise and clutter. The clutter-to-noise-ratio (CNR) is $55dB$. In the psd of the reference data, there is a jamming signal located at zero spatial frequency which is distributed over all Doppler frequencies. The jammer-to-noise-ratio is $40dB$. We conducted two groups of tests with schemes and parameters set exactly the same as in the tests for the simple model. The probability of detection of all these schemes, is shown in Fig. 14 as a function of SINR.

4.3 Discussion of Results

The results indicate that none of the schemes always outperforms all the others. Generally, SMI is not good for a case with mismatch between the reference data and the cell-under-test. If the additional hump is only in the cell-under-test psd then the additional hump is usually not suppressed by the processing. This may be the more common mismatch case. In the other case, where mismatch is due to an additional hump in the reference data psd, things are more complicated. In Fig. 11 and Fig. 13, SMI outperforms all the other schemes. A possible explanation is that, in the cases of Fig. 11 and Fig. 13, the reference data has statistics which are a good match to the data from the cell-under-test for angle and Doppler frequencies near the target. Since the effect of an extra hump is to overly suppress the signal returns from the specific Doppler frequency and spatial frequency where the mismatch located, this won't hurt too much if the target is far from that location. When there is no mismatch between the psd of the reference data and the psd of the data in cell-under-test, SMI is the optimal scheme. This also explained why SMI's performance was degraded in Fig. 12. In this case, the target's Doppler frequency is near the mismatch. We showed one case where the mismatch was due to an additional jamming signal in the reference data. In this case, ADPCA performed well.

ADPCA with overlapped pulses performs well in the most of the cases considered. This is especially true in those comparisons with element-space schemes. Although ADPCA has a similar structure to the element-space pre-Doppler schemes, it usually outperforms them. The reason is apparently based on the pulse canceling structure embedded in its steering vector, which makes it possible to cancel clutter with high correlation across several pulses even if this correlation is not present in the training data. ADPCA without overlapped pulses is not as good as ADPCA with the overlapped pulses. However, in most the cases, the performance difference between the two was not too large. Considering the computation saved, ADPCA without overlapped pulses may still be a good choice in practice. In the results in Fig. 9 (b), ADPCA does not perform very well, but in this case none of the pre-Doppler schemes perform very well.

5 Performance Comparison Using Measured Data

The multichannel airborne radar measurements (MCARM) program of Rome Laboratory is aimed at accelerating the development of STAP technology through the use of a common set of data. This program provides a database of measured airborne radar data which was collected by Westinghouse during several Delmarva and east coast fly-overs. Data used in this section comes from MCARM database flight 5 acquisition 575. Detailed information on the MCARM program and the measurements is available in [10] and [16]. Acquisition 575 includes non-homogeneous clutter. An animation of power spectrum of flight 5 acquisition 575 can be found at http://sunrise.oc.rl.af.mil/java/index.html. We inserted synthetic moving targets into different range bins to compare the detection performance of several STAP algorithms. This performance comparison includes ADPCA, factored post Doppler STAP, EFA and JDL. Reference data are selected from consecutive range cells on each side of the cell-under-test excluding the two closest guard cells. In the ADPCA implementation, each sub-CPI includes 3 consecutive pulses. Further, consecutive sub-CPIs overlap two pulses as shown in Fig. 6 (b). In the EFA scheme, adaptive processing is applied to 3 adjacent Doppler bins. In the JDL scheme, we define the LPR as a 3×3 square. See Appendix A for a summary of the parameters chosen for each algorithm tested.

In the first example, we inserted the synthetic target which corresponds to range bin 290 and Doppler bin 10. We employ a normalized test statistic (as in [14]) which provides a constant false alarm rate (CFAR) characteristic for homogeneous clutter. Results, in the form of the final test statistic, are provided for different range bins but these results are restricted to Doppler bin 10. In the results which appear in Fig. 15, we set the amount of reference data Q to 3 times the data vector length.

To evaluate the impact of the amount of reference data on detection performance, we reduce the amount of reference data Q to twice the data vector length. After the modification, the performance of the different schemes is shown in Fig. 16. We further reduce the amount of reference data Q to be equal to the data vector length. We show the performance in Fig. 17 for this case.

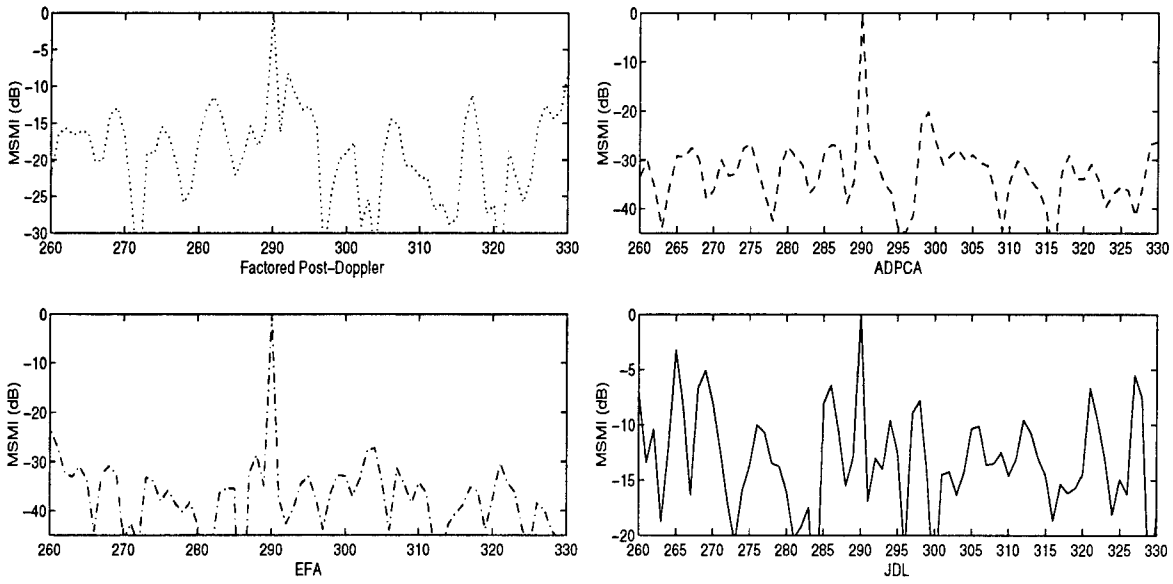


Figure 15: Performance comparison for example 1 with Q 3 times the data vector length.

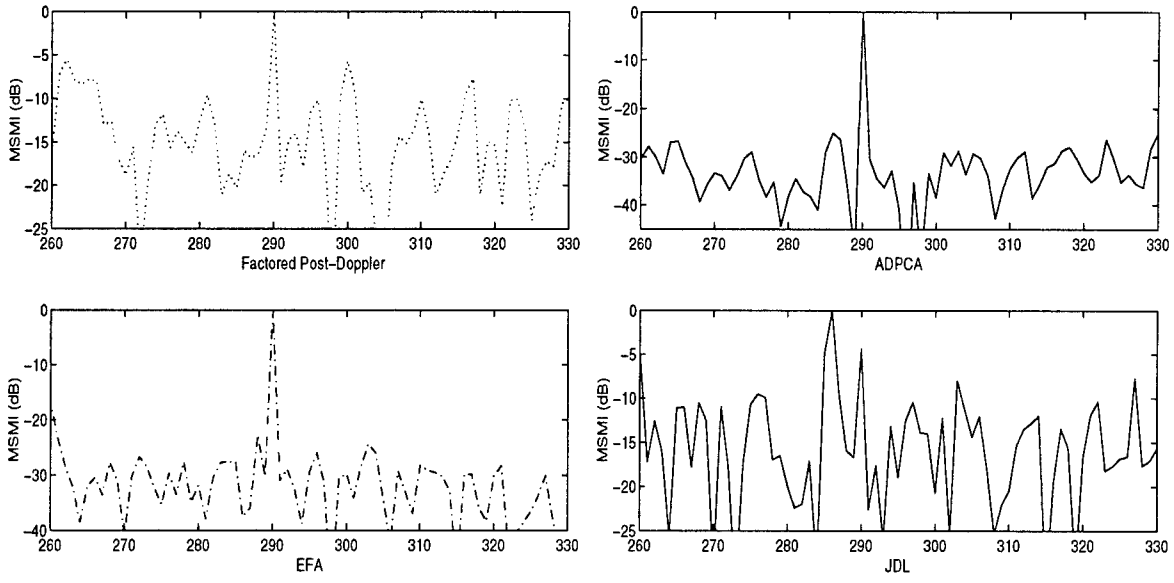


Figure 16: Performance comparison for example 1 with Q 2 times the data vector length.

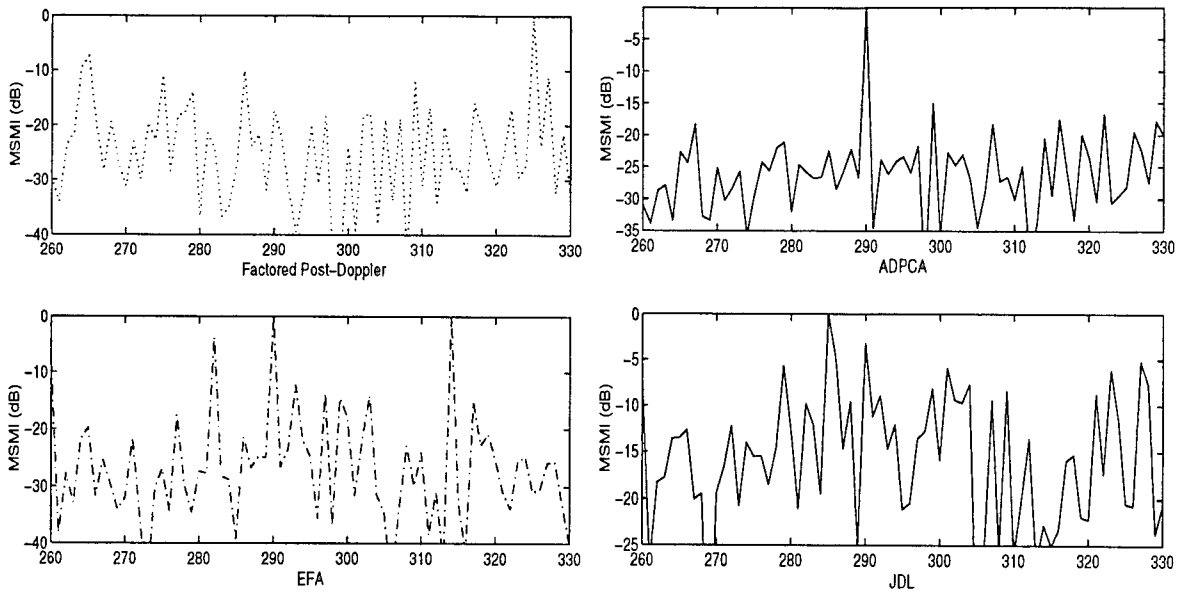


Figure 17: Performance comparison for example 1 with Q same as the data vector length.

In the second example, the synthetic target is inserted which corresponds to range bin 415 and Doppler bin 10. Similarly, we provide Fig. 18 through Fig. 20, which illustrate performance for various Q.

In measured data cases, it is hard to evaluate exactly how the statistics of the reference data and the statistics for the data from the cell-under-test are mismatched. A rough idea can be obtained from considering the energy fluctuation which occurs over range. In Fig. 21, we plot the energy in range bin 150 through 400 at Doppler bin 10 with the target signal in example 1. Here the target is located at range bin 290. In Fig. 22, we plot the energy in range bin 300 through 550 at Doppler bin 10 with the target signal in example 2. The target in example 2 is located at range bin 415. In Fig. 22, there is extreme variation in energy which includes step changes and linear variation in clutter power.

We note that ADPCA appears significantly more robust than EFA, JDL and factored post-Doppler in the results for example 1. Note that ADPCA is not affected significantly by the amount of reference data used. After reducing the amount of reference data used, ADPCA still performs well. The performance of all the other schemes degraded quickly as Q was reduced. To further explore ADPCA's potential, we reduced the amount of com-

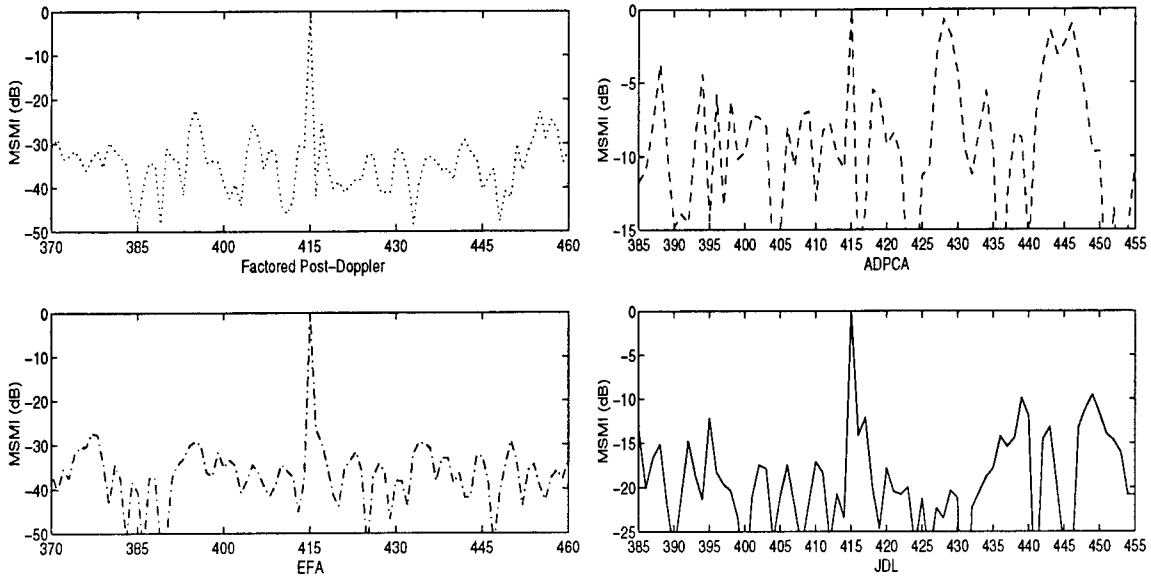


Figure 18: Performance comparison for example 2 with Q 3 times the data vector length.

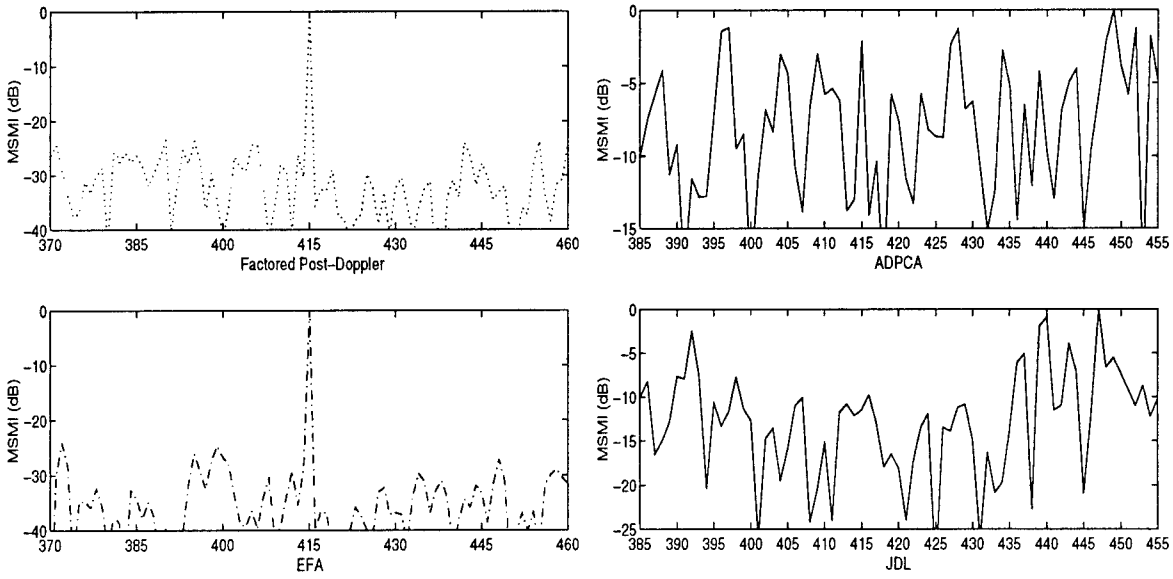


Figure 19: Performance comparison for example 2 with Q 2 times the data vector length.

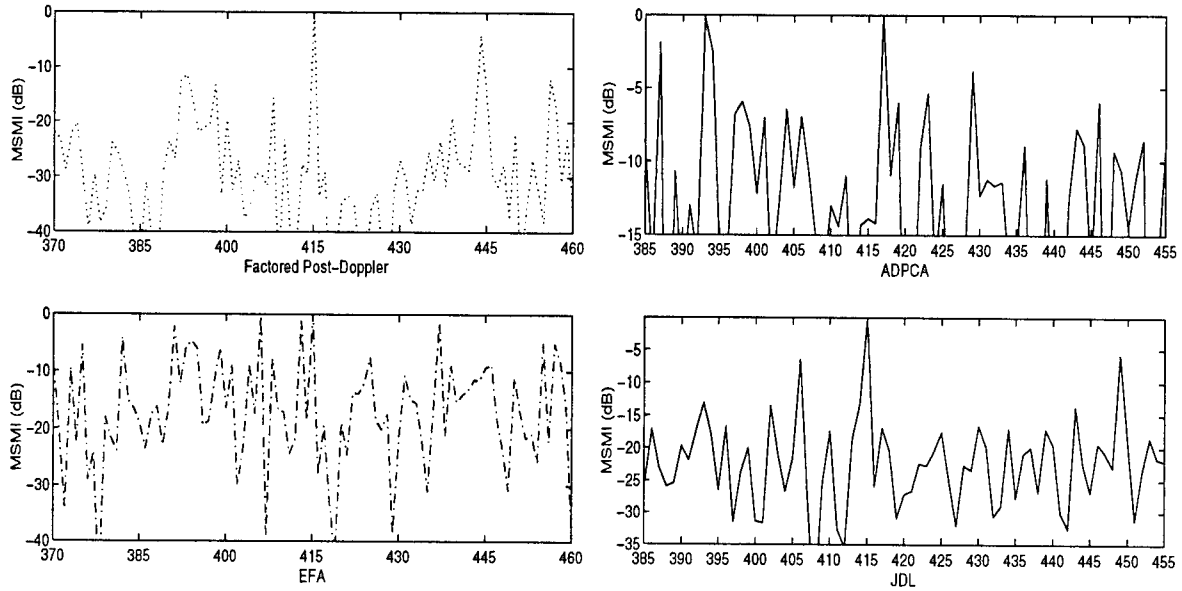


Figure 20: Performance comparison for example 2 with Q same as the data vector length.

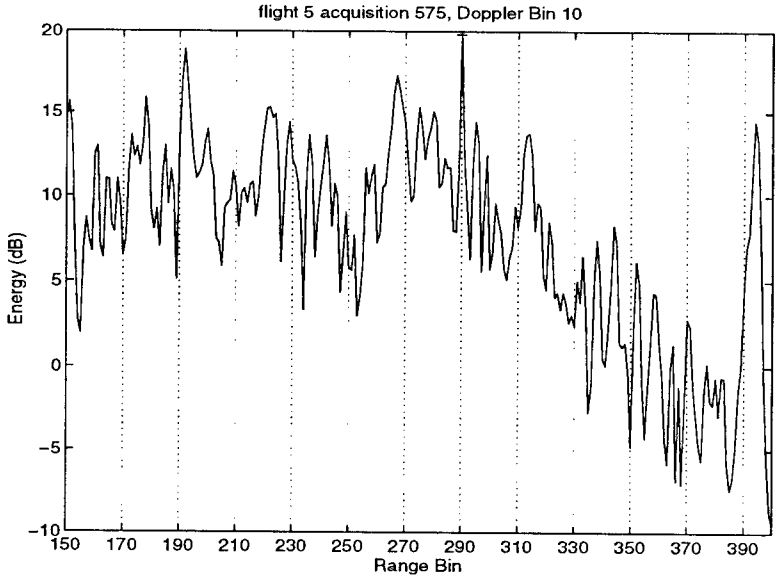


Figure 21: Energy for Doppler bin 10 and range bin 150 through 400 (target at range bin 290).

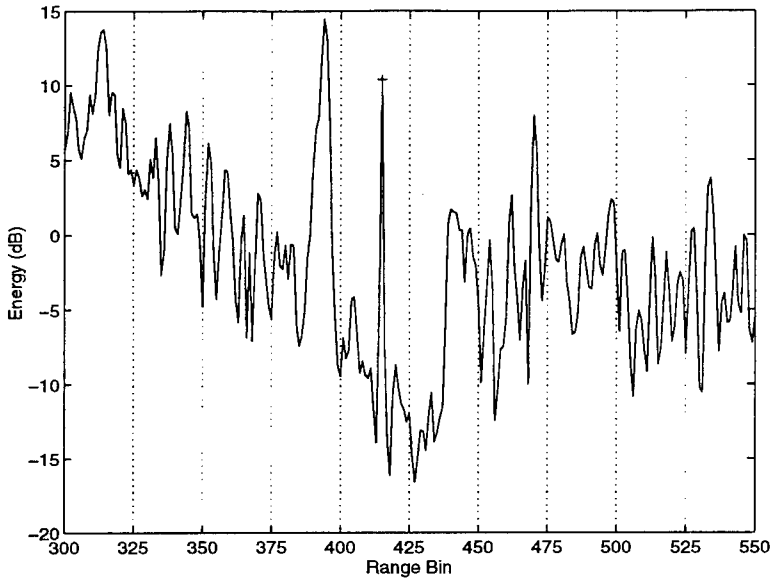


Figure 22: Energy for Doppler bin 10 and range bin 300 through 550 (target at range bin 415).

putations needed in ADPCA by estimating only one covariance matrix per range bin. We used the covariance matrix estimated for the first sub-CPI for all of the rest of the sub-CPIs corresponding to the same range bin. In this test, we set Q to 3 times the data vector length. As shown in Fig. 23 ADPCA still performs very well in this example.

However, in example 2, ADPCA fails to provide a distinguishable difference between the magnitude of the output at the target bin and the next highest competing clutter peak. Some of the other schemes did work well when there is enough quality reference data. Factored post-Doppler and EFA work well in this example. As in example 1, we tried testing ADPCA's performance when the computation is reduced by estimating one covariance matrix per range bin. Fig. 24 illustrates the results. To our surprise, these results are better than the results we obtained before we reduced the amount of computation. Similarly, we find that in Fig. 19 and Fig. 20 JDL's performance is improved after the amount of reference data is reduced. Reducing the amount of reference data used will not always cause degradation in performance in a severely non-homogeneous clutter environment. This is because the number of homogeneous range samples available for covariance estimation is limited. Thus,

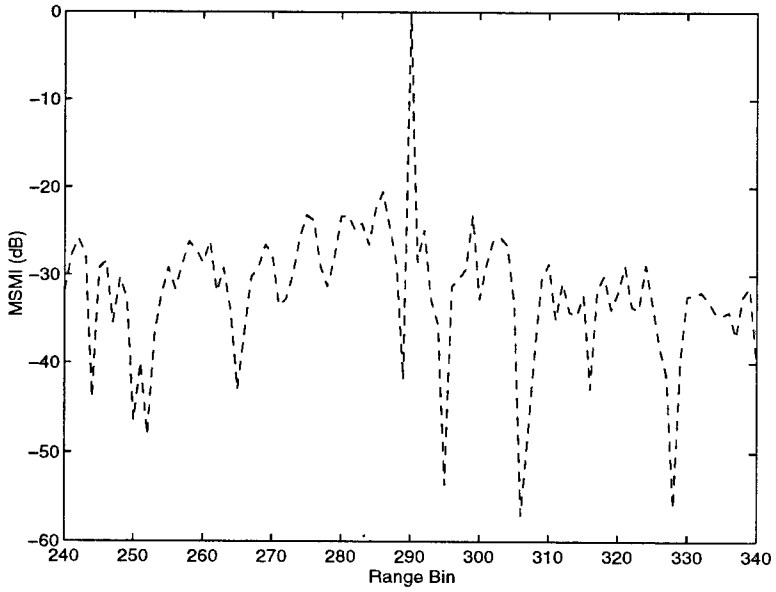


Figure 23: ADPCA with reduced complexity for the case where the target at Doppler bin 290.

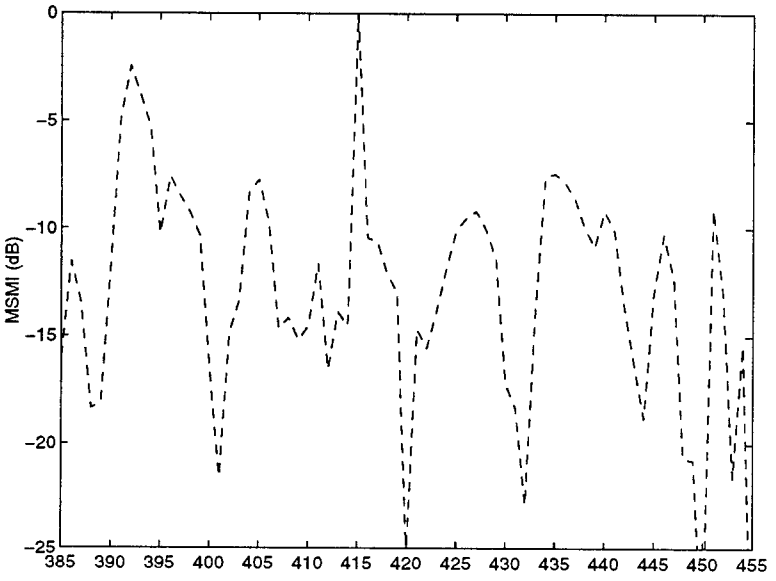


Figure 24: ADPCA with reduced complexity for the case where the target at Doppler bin 415.

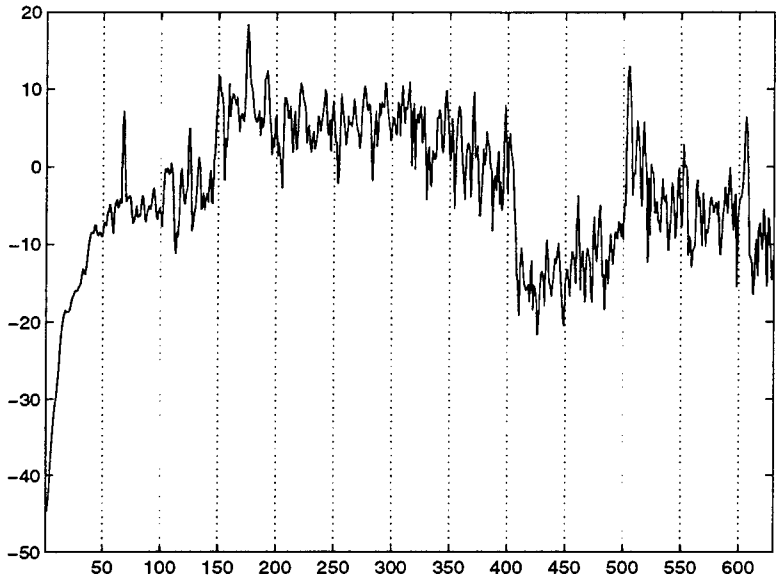


Figure 25: Energy for Doppler bin 30 and range bin 1 through 625 (target at range bin 150 of amplitude 0.1 is hard to see).

increasing the amount of reference data used could cause degradation. Actually, it is also the quality of the reference data, not just the amount of the reference data which determines the accuracy of the covariance estimation. In a severely non-homogeneous clutter environment, the situation could be subtle, a small part of the reference data could be dramatically different from the other part and it could distort the estimation. JDL uses the least amount of reference data of all the four schemes under comparison and it seems it is also the one most affected by the change in the amount of reference data. ADPCA does not work well in this example and we believe the reason is related to the step changes in energy near range bin 435 as shown in Fig. 22. The normalization of the test statistic which was imposed to achieve CFAR may also have contributed partially to the poor performance of ADPCA and JDL.

In the third example, a synthetic target is inserted which corresponds to range bin 150 and Doppler bin 30. We show the energy from range bin 1 through 625 in Doppler bin 30 in Fig. 25. Fig. 26 through Fig. 28 illustrate performance for various Q for example 3.

In example 3, ADPCA is not influenced much by changing the amount of reference

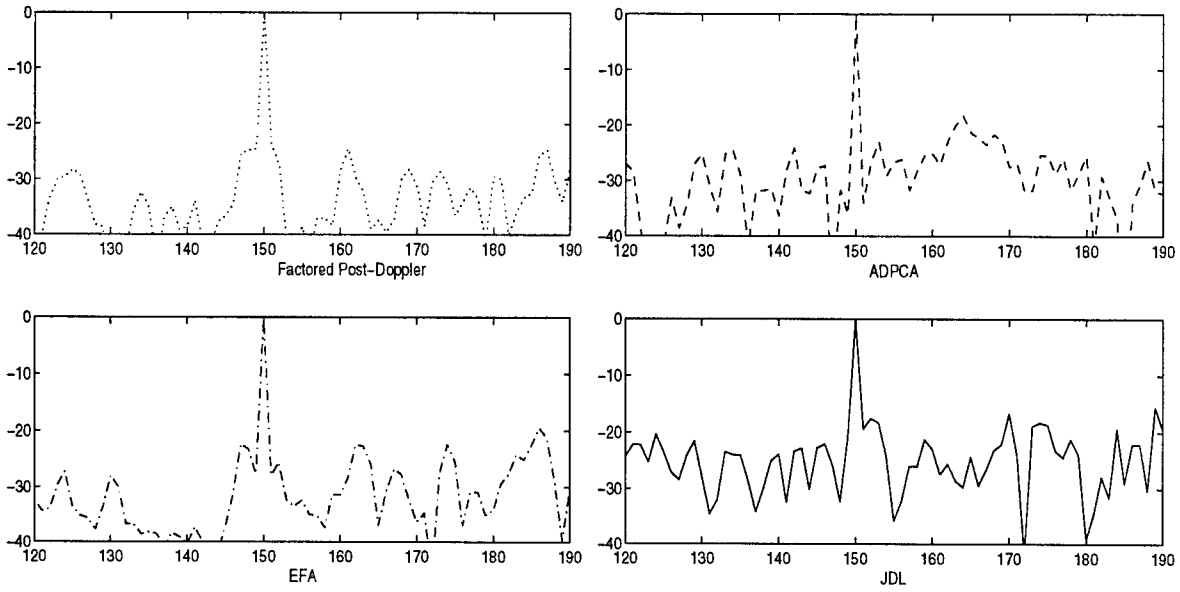


Figure 26: Performance comparison for example 3 with Q 3 times the data vector length.

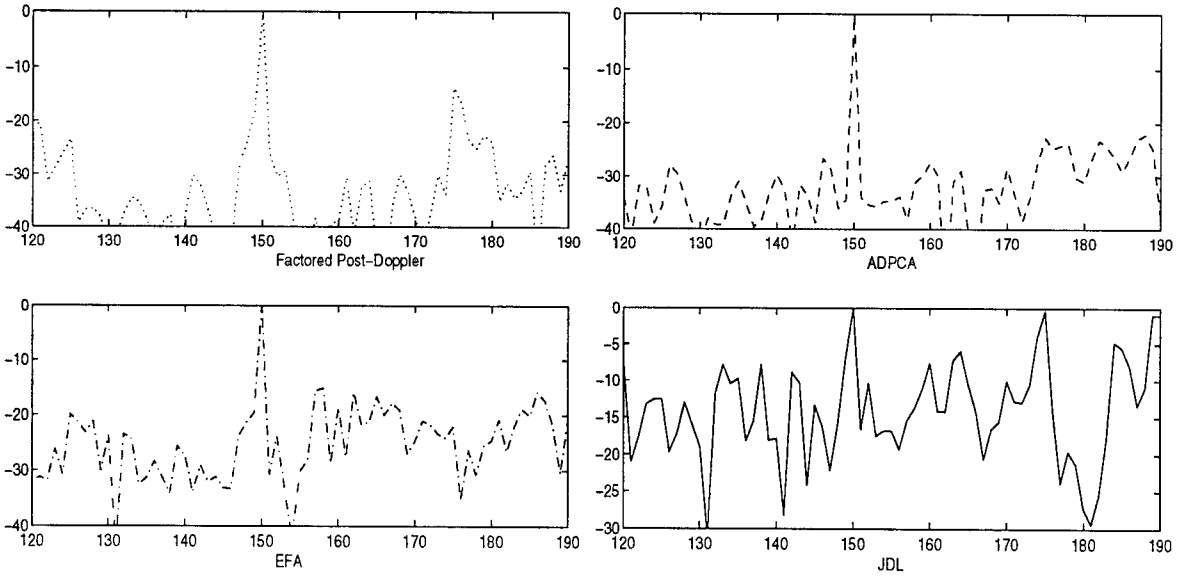


Figure 27: Performance comparison for example 3 with Q 2 times the data vector length.

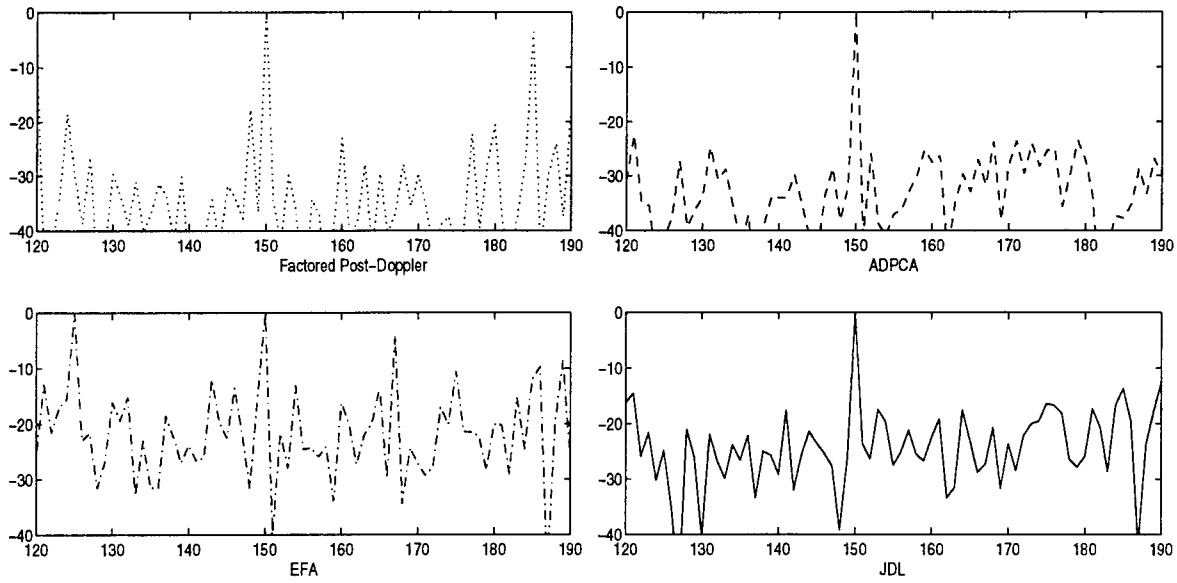


Figure 28: Performance comparison for example 3 with Q same as the data vector length.

data used. In Fig. 26, when the reference data size is 3 times the data vector length, factored post Doppler provided the biggest amplitude difference between the range bin in which target is located and the amplitude of the next highest peak. After reducing the amount of the reference data used, the performance of factored post Doppler degrades. JDL's performance seems sensitive to the quality of the reference data, as we already observed in the previous examples. An increase in the amount of reference data from one to two times the data vector length causes performance to degrade for JDL.

6 Conclusion

Our results show that the performance of SMI can be severely degraded in a non-homogeneous environment. This is especially true for cases where the reference data used in the estimation of interference-plus-noise statistics does not have statistics which closely match those for the data in the cell-under-test. However, a few STAP schemes appear to outperform SMI in these cases.

ADPCA is one of the promising schemes which performs well in many of the cases

studied in this report. However, there are some cases where ADPCA performs poorly. Using the knowledge we gained in the current research, we intend to develop generalizations of ADPCA, which will provide even better performance in a practical non-homogeneous clutter environment. One possible approach for developing these generalized schemes is to search for optimal pre-processing and post-processing transformations for cases with inaccurate parameter estimations. Modifying the steering vector so that it can work most effectively with theses transformations would also be interesting to study.

References

1. L. E. Brennan and I. S. Reed, "Theory of Adaptive Radar", *IEEE Trans. on Aerospace and Electronic Systems*, vol. AES-9, no. 2, pp. 237-252, March 1973.

2. I.S.Reed, J.D Mallett and L.E.Brennan, "Rapid convergence rate in adaptive arrays", *IEEE Transactions on Aerospace and Electronic Systems*, AES-10, no.6, November 1974.

3. Alfonso Farina, *Antenna-based Signal Processing Techniques for Radar Systems*, (Artech House: MA, 1992).

4. H. Wang, Y. Zhang and Q. Zhang, "A view of current status of space-time processing algorithm research", *IEEE International Radar Conference*, Alexandria, Virginia, May 1995, pp. 635-640.

5. D. Curtis Schleher, *MTI and Pulsed Doppler Radar*, (Artech House: MA, 1991).

6. Ramon Nitzberg, *Adaptive Signal Processing for Radar*, (Artech House: MA, 1992).

7. Fred E. Nathanson, *Radar Design Principles*, (McGraw-Hill Book Company: New York, 1969).

8. J. Ward, *Space-Time Adaptive Processing for Airborne Radar*, Technical Report 1015, Lincoln Laboratory, 1995.

9. E. C. Barile, R. L. Fante and J. Torres, " Some Limitations on the Effectiveness of Airborne Adaptive Radar", *IEEE Transaction on Aerospace and Electronic Systems*, vol. 28, No. 4, October 1992.

10. D. Sloper et.al., *MCARM Final Rept.*, Rome Lab Tech. Rept., RL-TR-96-49, April 1996.

11. R. C. DiPietro, "Extended Factored Space-Time Processing for Airborne Radar", *Proceedings of the 2th Asilomar Conference*, Pacific Grove CA, October 1992, pp. 425-430.

12. H. Wang and L. Cai, "On adaptive spatial-temporal processing for airborne surveillance radar systems", *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 30, pp. 660-669, July 1994.

13. I. R. Roman and D.W. Davis, *Multichannel System Identification and Detection using output Data Techniques*, Final Report Report No. SSC-TR-96-02.

14. W. L. Melvin, C. Wicks and R. D. Brown, "Assessment of Multichannel Airborne Radar Measurements for Analysis and Design of Space-Time Processing Architectures and Algorithms", *Proceedings of the 1996 IEEE National Radar Conference*, Ann Arbor, MI, May 1996, pp. 130-135.

15. H. Wang, H.R. Pard and M.C. Wicks, "Recent Results in Space-Time Processing", *IEEE International Radar Conference*, Atlanta, Georgia, March 1994, pp. 104-109.

16. W. L. Melvin and B. Himed, "Comparative Analysis of Space-Time Adaptive Algorithms with Measured Airborne Data", *presented at the 7th International Conference on Signal Processing Applications and Technology*, October 7-10, 1996.

Appendix

Most of the parameters used in the comparisons are listed in the following tables. Here, we use M to denote the number of pulses per CPI, N to denote the number of elements. Definitions of A_p, B_p and f_m can be found in (9) and (13) where pre-processing and post-processing are defined. We define F_K as a $K \times K$ DFT matrix and $f_{K,p}$ is its pth column. I_r stands for an $r \times r$ identity matrix. The notation $0_{l \times m}$ refers to an $l \times m$ matrix of zeros. Dp is used to denote the Doppler bin where target is located. NA stands for Not Applicable.

Scheme	M	N	p	A_p	B_p	f_m
SMI	12	2	0	I_{12}	I_2	NA
ADPCA (overlap)	12	2	0...9	$\begin{bmatrix} 0_{p\times 3} \\ I_3 \\ 0_{(9-p)\times 3} \end{bmatrix}$	I_2	$f_{10,m}$
ADPCA (without overlap)	12	2	0...3	$\begin{bmatrix} 0_{3p\times 3} \\ I_3 \\ 0_{(9-3p)\times 3} \end{bmatrix}$	I_2	$f_{4,m}$
Element-Space Pre-Doppler	12	2	0...9	$\begin{bmatrix} 0_{p\times 3} \\ I_3 \\ 0_{(9-p)\times 3} \end{bmatrix}$	I_2	$f_{10,m}$
Factored Post-Doppler	12	2	0	$f_{12,Dp}$	I_2	NA

Table 2: Parameters for comparison tests in Section 4 group (a).

Scheme	M	N	p	A_p	B_p	f_m
SMI	12	4	0	I_{12}	I_4	NA
ADPCA (overlap)	12	4	0...9	$\begin{bmatrix} 0_{p\times 3} \\ I_3 \\ 0_{(9-p)\times 3} \end{bmatrix}$	I_4	$f_{10,Dp}$
ADPCA (without overlap)	12	4	0...3	$\begin{bmatrix} 0_{3p\times 3} \\ I_3 \\ 0_{(9-3p)\times 3} \end{bmatrix}$	I_4	$f_{4,Dp}$
Beam-Space Pre-Doppler	12	4	0...9	$\begin{bmatrix} 0_{p\times 3} \\ I_3 \\ 0_{(9-p)\times 3} \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}$	$f_{10,Dp}$
Beam-Space Post-Doppler	12	4	0	$f_{12,Dp}$	$\begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}$	NA

Table 3: Parameters for comparison tests in Section 4 group (b).

Scheme	M	N	p	A_p	B_p	f_m
Factored Post-Doppler	128	22	0	$f_{128,Dp}$	I_{22}	NA
EFA	128	22	0	$[f_{128,Dp-1}, f_{128,Dp}, f_{128,Dp+1}]$	I_{22}	NA
ADPCA	128	22	0...126	$\begin{bmatrix} 0_{p\times 3} \\ I_3 \\ 0_{(125-p)\times 3} \end{bmatrix}$	I_{22}	$f_{126,Dp}$
JDL	128	22	0	$[f_{128,Dp-1}, f_{128,Dp}, f_{128,Dp+1}]$	*	NA

Table 4: Parameters for comparison tests in Section 5.

*MCARM data is not collected by a uniformly spaced linear antenna array, so its beamforming matrix is relatively complicated. This matrix is provided with the MCARM database.

Mostafa Chinichian
Report not available at time of publication.

DEVELOPMENT OF EFFICIENT ALGORITHMS AND SOFTWARE CODES FOR LOSSLESS AND NEAR-LOSSLESS COMPRESSION OF DIGITIZED IMAGES

Manohar K. Das
Associate Professor
Department of Electrical and Systems Engineering

Oakland University
Rochester, MI 48309-4401

Final Report for:
Summer Research Extension Program

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington DC

and

Oakland University

December 1996

Development of Efficient Algorithms and Software Codes for Lossless and Near-Lossless Compression of Digitized Images

Abstract

As modern digital imaging technology makes its ever-growing inroads in today's military and commercial applications, computer scientists and engineers face the increasingly difficult task of archiving, transmitting, and manipulating gigantic volumes of image data that result from the process of digitization. Although the cost of storing and transmitting data continues to decrease, such reductions seem only to stimulate new applications encompassing ever-increasing use of digital information. An excellent solution to this data explosion problem is afforded by a lossless or near-lossless image compression scheme, which will significantly reduce the cost of storage or transmission, and at the same time, allow subsequent recovery of the images with either very little, or no loss at all. The goal of this project is to investigate such a solution.

Based primarily on the PI's past experience and the outcome of his research conducted as a Summer Faculty Research Participant at Rome Laboratory, New York, three specific objectives are addressed in this project; namely, i) development of usable software codes for two new image compression algorithms developed by the PI during his summer research at Rome Laboratory, ii) further improvement of their compression performance by incorporating context-dependent source coders, and iii) development of an efficient scheme for coding and transmission of thumbnail images, used at Rome Laboratory and elsewhere within the Air Force.

The results of the experimental studies indicate that the proposed coding schemes hold much promise for widespread applications in transmission and archival of images used by the Air Force.

1. INTRODUCTION

The sheer explosion of data resulting from the digitization process poses a formidable problem for widespread use of digital techniques for transmission and/or archival of raw images. For instance, a 2048x2048 image, quantized to 8 bits per pixel, requires storage and/or transmission of approximately 4.2 megabytes of picture data. The lossless transmission of such an image over a 19.2 kilobits/second channel requires approximately 29 minutes, which is rather long. The problem becomes even more acute for video pictures, such as, color television scenes. Typical television images have a spatial resolution of approximately 512x512 per frame. Assuming 30 frames are transmitted per second and 24 bits are used to digitize the intensity and color information of each pixel, the storage of only one second of the signal would require approximately 23.6 megabytes. Transmitting such a signal in real time would require a 188.7 million bits/second channel.

One of the most promising solutions to the above data explosion problems is afforded by image coding techniques. The goal of such a technique is concerned with the reduction of the number of bits required to store or transmit images with either little or no loss of any information. For instance, an information preserving image coding technique, which delivers a bit rate of the order of 2 bits/pixel, will reduce an original 8 bits/pixel image file to only 20% of its original size, with a concomitant four fold reduction in the transmission time.

Image data compression has been a popular research topic for the past two decades. As a result of the concerted efforts of several researchers, a multitude of different image coding techniques have emerged [1]-[4], most of which can be broadly categorized into three classes; namely, lossy, lossless, and near-lossless. A lossy scheme [5]-[12] does not allow exact recovery of the original image, but can deliver high compression ratios, e.g., fifty or more; a lossless scheme [13]-[24], on the other hand, allows exact recovery of the original image, but delivers a compression of the order 3 or 4 only. Finally, a near-lossless scheme [25]-[27] can deliver significantly higher compression compared to the lossless ones, and at the same time, guarantee that the reconstruction error for each pixel will lie within some predefined limits chosen by the user. For instance, the near-lossless compression of an 8-bit (i.e., 256 gray levels) image can typically result in an eight fold reduction of the original file size, while still guaranteeing that the reconstruction error for each pixel will lie within say, ± 1, or ± 2 of the original gray level value.

This research project was conceived as an extension of the summer research program pursued by the PI at Rome Laboratory, New York, in the summer of 1995. The project focuses on development of software codes for two new lossless and near-lossless image coders investigated by the PI at Rome Laboratory, and study of some novel image coding algorithms. The specific objectives are summarized below.

2. OBJECTIVES OF THIS PROJECT

The three major objectives of this project are:

i. ***Development of usable "C" codes for two new image coding algorithms developed by the PI during his summer research at Rome Laboratory***

This involves: a) development of "C" codes for two novel image coding algorithms called suboptimal adaptive DPCM (SADPCM) [27] and hierarchical block-adaptive DPCM (HBADPCM) [28], b) generation of efficient "C" codes for compression and expansion modules, and c) testing of the codes.

ii. ***Development of new, improved lossless/near-lossless image coders by combining SADPCM and HBADPCM with context dependent source coders***

This involves: a) selection of an appropriate image modeler that offers the best trade-off between cost and performance, b) selection of a good contextual source coder [29]-[31], and c) experimentation with a variety of images.

iii. ***Development of an efficient coding scheme for storage and/or transmission of thumbnail images used at Rome Laboratory***

This involves: a) selection of an edge preserving filter for smoothing of thumbnail images, followed by coding using either lossy SADPCM or JPEG coders, b) selection of the most appropriate multiresolution decomposition scheme based on a comparative evaluation of the available techniques, such as, Laplacian pyramid method [8], difference pyramid technique [19],[32],[33], and reduced difference pyramid decomposition [20], and c) selection of the most appropriate coding strategy, such as, DPCM coders, 2-D MAR coders [12], or JPEG [3].

The following sections provide detailed discussion of the methodologies used to realize the above objectives, and experimental results. Specifically, Section 3 describes the usable "C" codes developed for the SADPCM and HBADPCM algorithms, whereas further improvements of these algorithms through incorporation of context-dependent entropy coders is addressed in Section 4. Also, the coding of thumbnail images is discussed in Section 5, and finally, some concluding remarks are given in Section 6.

3. DEVELOPMENT OF USABLE "C" CODES FOR SADPCM AND HBADPCM ALGORITHMS

Two efficient algorithms for lossless compression of digitized images were developed by the PI during his tenure as a Summer Research Fellow at Rome Laboratory, New York. These represent adaptive and hierarchical versions of the conventional fixed differential pulse code modulation (DPCM) scheme,

and are called suboptimal adaptive DPCM (SADPCM) [27] and hierarchical block-adaptive DPCM (HBADPCM) [28], respectively. The SADPCM algorithm is designed for full-frame compression of monochrome still images, and its main advantages include: i) easy implementation, ii) fast execution, iii) higher lossless/near-lossless compression compared to conventional fixed DPCM method (by a margin of about 8% or more). On the other hand, the HBADPCM algorithm is designed for progressive or hierarchical compression, where a coarse reproduction of the image is first constructed using a small fraction of the pixels, and this coarse reproduction is gradually updated as the remaining pixels arrive. Progressive coding is particularly helpful when browsing a large image data base, because in this case the user can quickly determine from the coarse version of the image whether it is an appropriate one. If not, the user can realize significant cost savings by stopping the transmission of the finer versions of the image. The main advantages of the proposed HBADPCM algorithm include: i) easy implementation, ii) fast execution, iii) higher lossless/near-lossless compression compared to conventional hierarchical interpolation (HINT) [13] scheme (by a margin of about 10% or more), and iv) selective, segment-by-segment transmission of an image frame.

Brief descriptions of the above algorithms are provided in the following subsections. Details can be found in the references cited above.

3.1 The SADPCM Image Modeling and Estimation Approach

To begin with, we assume that a two-dimensional (2-D) digitized image can be regarded as a nonstationary 2-D signal consisting of pixel intensity values, $\{f(i,j), 1 \le i \le L, 1 \le j \le L\}$, where i denotes the row index and j stands for the column index, respectively. A conventional 2-D DPCM compression scheme uses an image model of the following form:

$$f(i,j) - a_1 f(i,j-1) - a_2 f(i-1,j) + a_1 a_2 f(i-1,j-1) = w(i,j), \tag{1}$$

where a_1 and a_2 denote the model coefficients, and w(i,j) denotes the modeling error, which is usually regarded as a zero-mean, 2-D white noise sequence. In case of conventional fixed DPCM schemes, it is customary to set $a_1 = a_2$ and a commonly used value for both coefficients is 0.95 [5]. Henceforth, this scheme is referred to as the fixed DPCM technique.

In order to develop a suboptimal adaptive DPCM (SADPCM) image model, first equn. (1) is rewritten in the form of a 2-D multiplicative autoregressive model [12],

$$(1 - a_2 q_1^{-1})(1 - a_1 q_2^{-1})\, f(i,j) = w(i,j), \tag{2}$$

where q_1^{-1} and q_2^{-1} denote the unit backward shift operators in the vertical (i.e., row-wise) and horizontal (i.e., column-wise) directions, respectively. Because of the multiplicative nature of the polynomial operator in the left side of (2), we can further express (2) in the cascade form,

$$(1 - a_1 q_2^{-1})\, f(i,j) = f_1(i,j), \tag{3a}$$

$$(1 - a_2 q_1^{-1})\, f_1(i,j) = w(i,j), \tag{3b}$$

where $f_1(i,j)$ is an intermediate signal regarded as the output of the first stage of the cascade structure.

In order to be useful for adaptive image coding, the coefficients a_1 and a_2, of model (3) must be estimated from the given image data. As discussed in [12], an optimal method of estimating a_1 and a_2 involves either minimization of a nonlinear cost function, or a pseudo-linearization approach. However, both of these approaches are beset with moderately high computational cost, and therefore, a suboptimal estimation scheme, as described below, is deemed to be more appealing from the point of view of practical implementation.

3.1.1 A suboptimal scheme for estimation of the model coefficients

The selection of a suboptimal estimation scheme depends on the trade-off between performance and computational complexity. The simplest coders, presented in this section, use a fixed value for a_1 and an estimated value for a_2, whereas for the improved coders, presented in a subsequent section, estimated values of both a_1 and a_2 are utilized. To find the estimates of both a_1 and a_2, the following three-step suboptimal strategy may be used:

Step 1. Estimate a_1 from (3a) ignoring the correlation, if any, between $f_1(i,j)$ and $f(i,j)$;

Step 2. Using the estimated value of a_1, evaluate the residuals of the first stage to form an estimate, $\hat{f}_1(i,j)$, of $f_1(i,j)$ as,

$$\hat{f}_1(i,j) = f(i,j) - a_1 f(i,j-1);$$

Step 3. Finally, estimate a_2 from (3b).

Notice that the above procedure is a suboptimal one because the estimate of a_1 is biased due to non-zero correlation between $f(i,j)$ and $f_1(i,j)$. However, it is still useful for predictive coding because of two reasons: i) in a predictive coder, the modeling inadequacies get masked to a certain extent through the process of adding back the quantized residual errors during reconstruction, and ii) the error cannot grow provided the estimated model is stable.

Using the above strategy, estimates of a_1 and a_2 are simply obtained as,

$$a_1 = r_f(0,1)/r_f(0,0), \tag{4a}$$

$$a_2 = r_{f1}(1,0)/r_{f1}(0,0), \tag{4b}$$

where $r_f(k,l)$ and $r_{f1}(k,l)$ denote the autocorrelation coefficients with 'lag (k,l)' of $f(i,j)$ and $f_1(i,j)$, respectively. Notice that for raw image data, the value of a_1 almost always lies between 0.95 and 1.0, which represents the typical range of normalized column-wise autocorrelation coefficients for most images. Thus, for the simplest SADPCM coders, which process $f(i,j)$ directly, the value of a_1 can be assumed to be a constant.

Next, notice that the estimation of a_2 can be carried out either in a forward, or backward fashion.

In the forward scheme, the image is first subdivided into smaller, say MxM, blocks, and a separate set of coefficients is estimated for each block. In this case, the blockwise estimates of a_2 need to be transmitted as the side information. On the other hand, in the backward scheme, the estimation of a_2 is carried out over a causal window consisting of a few reconstructed pixel values, and therefore, avoids the necessity of transmitting any side information. The overall coding schemes based on forward and backward estimation techniques are henceforth referred to as forward ADPCM (FADPCM) and backward ADPCM (BADPCM), respectively.

In the forward scheme, two alternative estimation strategies are useful; namely, batch and recursive methods. The batch method involves blockwise estimation of $r_{f1}(1,0)$ and $r_{f1}(0,0)$, followed by evaluation of a_2 from (4b). The recursive method, on the other hand, involves calculation of a_2 using a recursive estimation algorithm, such as, recursive least squares (RLS) or least mean squares (LMS) [9]. Since we need to estimate a single parameter, the use of RLS is preferred because it converges faster and involves only a little extra computation than LMS. The RLS estimation update equations can be summarized as follows:

error update: $$e(i,j) = f_1(i,j) - a_2 f_1(i-1,j), \tag{5a}$$

covariance update: $$p^u = p^o/g, \tag{5b}$$

normalizing gain: $$g = 1 + p^o[f_1(i,j)]^2, \tag{5c}$$

parameter update: $$a_2{}^u = a_2{}^o + p^u f_1(i,j)e(i,j), \tag{5d}$$

where the subscripts 'o' and 'u' refer to the old and the updated values, respectively, of the associated variables.

In the backward scheme, the use of a recursive estimation technique is mandated, and two alternative strategies are found to be useful. The first method uses either RLS or LMS to update the a_2 estimate for each pixel. In the second method, first the values of $r_{f1}(1,0)$ and $r_{f1}(0,0)$ are updated over a sliding causal window, and then pixel-by-pixel estimates of a_2 are calculated from (4b). The window used in some of our experiments is shown in Fig. 1. Using this window, the estimated values of $r_{f1}(0,0)$ and $r_{f1}(1,0)$ for the $(i,j)^{th}$ pixel are obtained as,

$$r_{f1}(0,0) = \Big(\sum_{m,n\in S} \sum [f_1(i-m,j-n)]^2\Big)/7.0, \tag{6a}$$

$$r_{f1}(1,0) = [f_1(i,j-1)f_1(i-1,j-1) + f_1(i-1,j-1)f_1(i-2,j-1) + f_1(i-1,j)f_1(i-2,j) + f_1(i-1,j+1)f_1(i-2,j+1)]/4.0, \tag{6b}$$

where $S=\{(m,n) \mid m\in[0,2],\ n\in[-1,1],\ n \neq (0 \text{ or } 1) \text{ for } m=0\}$. It may be mentioned that because of the partial overlap of the windows for two neighboring pixels, the actual computational load required for

implementation of (6a) and (6b) is only about five multiplications and four additions per pixel, which can be reduced further by choosing a window of smaller size.

In both forward and backward approaches, the robustness of the overall scheme is greatly improved by adding a stability check for $\hat{a}_2$, the estimated value of a_2, at every step. Notice that the stability of the estimated model is guaranteed if

$$|\hat{a}_2| < 1. \tag{7}$$

If (7) is violated, the estimate is first projected inside the stable zone before continuing with prediction and encoding steps.

Next, lossless and near-lossless image coding schemes based on the above SADPCM modeling approach is presented below.

3.2 Lossless Image Coding Using SADPCM Modeling Approach

As mentioned earlier, two different coding schemes are possible; namely, FADPCM and BADPCM. Since they differ only in the estimation methodology, we summarize the overall coding scheme below for a generic SADPCM model only.

A lossless image coder based on the SADPCM model consists of two main components, namely, a predictor and an entropy coder. The predictor simply uses (1) to calculate a predicted value of f(i,j) as,

$$\tilde{f}(i,j) = 0.99f(i,j-1) + \hat{a}_2 f(i-1,j) - 0.99\hat{a}_2 f(i-1,j-1), \tag{8}$$

where a_1 is chosen to be 0.99 and $\hat{a}_2$ denotes the estimated value of a_2. Next, $\tilde{f}(i,j)$ is rounded to generate the integer predicted values, $f_r(i,j)$, i.e.,

$$f_r(i,j) = R[\tilde{f}(i,j)], \tag{9}$$

where R[x] denotes the nearest integer value of x. The residual signals, d(i,j), are then obtained as,

$$d(i,j) = f(i,j) - f_r(i,j). \tag{10}$$

Finally, the residual sequence, {d(i,j)}, is entropy coded using an optimal encoder and the coded residuals are transmitted to the receiver. At the receiver end, f(i,j) is exactly reconstructed by first synthesizing $f_r(i,j)$ using (8), (9) and then utilizing the decoded d(i,j) to obtain

$$f(i,j) = f_r(i,j) + d(i,j). \tag{11}$$

3.3 Improved SADPCM Coding Schemes

It turns out that the coding performance of the above DPCM coders can be improved significantly with a little additional computational cost. The basic idea and the motivation behind it are summarized below.

Notice that most image data are nonstationary-in-the-mean, which is usually caused by the

combined effect of both reflectance and lighting variations over different parts of an object or scene. Since predictive coders work best on stationary data, it makes sense to attempt to improve the performance of the SADPCM coders by removing the nonstationary attributes of the image data.

The simplest method of eliminating nonstationarity-in-the-mean consists of subtraction of an estimated local mean from each pixel value. Since the SADPCM coders utilize two coefficients associated with the nearest causal neighbors of each pixel, f(i,j), it makes sense to obtain the staionary-in-the-mean image as follows:

$$f_m(i,j) = R[0.5*(f(i,j-1) + f(i-1,j))], \quad (16a)$$

$$f_s(i,j) = f(i,j) - f_m(i,j), \quad (16b)$$

where $f_m(i,j)$ denotes the local mean at location (i,j), R[.] denotes the operation of rounding to the nearest integer value, and $\{f_s(i,j)\}$ denotes the stationary-in-the-mean image data.

Next, the SADPCM scheme is utilized to code $\{f_s(i,j)\}$ using either a forward or a backward approach, as discussed earlier. However, in this case, both a_1 and a_2 need to be estimated because a fixed value for a_1 cannot be assumed any more. The modified methods are henceforth called improved FADPCM (I-FADPCM) and improved BADPCM (I-BADPCM) schemes, respectively.

3.4 Lossless HBADPCM coders

There are two main reasons that can be regarded as the motivating factors for pursuing the development of hierarchical coders. First, it is desirable to adapt SADPCM coders for hierarchical transmission in order to achieve better compression than that afforded by the existing simple and elegant coders, such as, HINT. Second, in many applications, it is desirable to develop a scheme that allows selective, segment-by-segment transmission of an image frame instead of one full frame at a time. As far as the first goal is concerned, since predictive coders work well only on a full image frame or subsections of it, and performs rather poorly on subsampled images, a block-by-block encoding and transmission method is attempted in the so called hierarchical block-adaptive DPCM (HBADPCM) method [28], developed by the PI at Rome Laboratory. This also seems to be good for meeting the second goal because an user can selectively access different segments of an image in any desired sequence. The implementation of HBADPCM can be carried out using two alternative approaches, namely, forward or backward, as discussed below.

3.4.1 A Forward Hierarchical Block-Adaptive DPCM (F-HBADPCM) Scheme

In order to achieve blockwise hierarchical transmission, first the image is subdivided into a number of smaller, say MxM, blocks. The next thing needed is adoption of a suitable hierarchical framework for coding and transmitting different blocks. Although a variety of different hierarchical frameworks could have been chosen for this purpose, that of HINT is selected mainly because of its excellent performance

as a hierarchical lossless coder. Also, as shown in the following subsection, this framework allows a nice implementation of a backward HBADPCM coder.

Following [13] then, we arrange the blocks to be coded in the hierarchical order illustrated in Fig. 2. In this figure, the blocks are marked according to their designated level of hierarchy. Thus, the blocks marked "4" denote the ones to be coded and transmitted first, followed by the blocks marked "3", and so on.

Next, the coding and transmission of each individual block is carried out using either FADPCM (which requires blockwise estimation of a_2 only), or its improved version, I-FADPCM (which requires blockwise estimation of both a_1 and a_2). Then the block-by-block residuals pertaining to each hierarchical level are encoded using an entropy coder and the coded residuals are transmitted. It is pointed out that there is no need for construction of separate codebooks for different hierarchical levels, because all the residuals generated by FADPCM or I-FADPCM exhibit narrow peaks and can be efficiently coded using a single codebook. The overall coding schemes are henceforth referred to as forward HBADPCM (F-HBADPCM) and improved F-HBADPCM, respectively.

Until now, our only use of the HINT framework has been restricted to the task of laying out a hierarchy of blocks. As expected, fruitful exploitation of the above framework allows further improvement of the HBADPCM coder discussed above. One such improvement, namely, a backward HBADPCM coding scheme, is presented below.

3.4.2 A Backward Hierarchical Block-Adaptive DPCM (B-HBADPCM) Scheme

The main objective of a backward HBADPCM scheme would be to modify the parameter estimation strategy such that transmission of the block-by-block model coefficients becomes unnecessary. To realize the above goal, two useful strategies have been tried: i) a backward estimation scheme, where the model coefficients pertaining to a block are estimated from the past reconstructed pixel values within that block, and ii) a hierarchical approximation scheme, where the block-by-block model coefficients are first laid out in the same hierarchical pattern as the blocks themselves, and an approximate set of blockwise model coefficients is formed through hierarchical interpolation from the past reconstructed blocks' coefficients. This set of approximate coefficients is used by both the transmitter and the receiver, avoiding the necessity of transmitting any side information.

A comparative experimental study of the two backward HBADPCM schemes indicates that the hierarchical approximation scheme performs slightly better than the backward approximation one. Because of this and the fact that the latter attempts a true exploitation of the HINT framework, the HINT approximation approach is adopted here. However, it may be pointed out that like in the forward case, there exist two versions of the scheme; namely, one based on blockwise estimates of a_2 only, and the other

based on blockwise estimates of both a_1 and a_2. The overall coding schemes are henceforth referred to as the backward HBADPCM (B-HBADPCM), and improved B-HBADPCM (IB-HBADPCM), respectively.

Next, one important thing that needs to be addressed relates to the question of how to interpolate for the missing blocks so that higher level approximations can be generated, if needed, from the reconstructed image blocks pertaining to a lower hierarchal level. This issue is addressed in the next subsection.

3.4.3 A Scheme for Interpolation for the Missing Blocks

Notice that the problem of interpolation for the missing blocks, in the above context, bears a striking similarity to the problem of error concealment in case of loss of cells during transmission [36], [37]. Thus, techniques used for error concealment - particularly, the spatial domain ones, are well suited to our task.

In view of above, a simpler version of the projective interpolation technique [37], is attempted here for demonstration of interpolation for the missing blocks. In the projective interpolation scheme, Jung et al suggests using bilinear interpolation adapted to the edge pattern of a missing block. Essentially, their method consists of four steps; namely, i) examination of the boundary pixel values of a missing block for possible edge patterns within the missing one, ii) classification of a missing block into one of six categories depending on the possible edge patterns, iii) determination of the interpolation direction, and iv) interpolation of the missing block using a weighted bilinear interpolation.

A simpler version of the above technique consists of the following procedure:

- For each pixel within a missing block, examine four pairs of boundary pixel values - located along the horizontal, vertical, 45^0, and 135^0 directions;
- Identify the pair possessing the minimum absolute difference among the four, and perform a linear interpolation in the corresponding direction.

3.5 Software Codes

After experimenting with a wide variety of test images, it was decided to code the I-FADPCM and F-HBADPCM algorithms, because they perform somewhat better than their backward counterparts. For the sake of simplicity, these two algorithms are simply referred to as SADPCM and HBADPCM, respectively, in the discussion to follow.

The software developed for each of the above algorithms consists of two basic modules: a compressor and an expander. The compressor module, in turn, consists of three sub-modules; namely, a residual generator, an integer-to-bit-stream converter, and a shifted Huffman encoder. The residual generator constructs integer residual values using one of the algorithms mentioned earlier, the integer-to-bit-stream converter converts the residuals first into a string of ASCII characters and then into a bit-stream,

and finally, the shifted Huffman encoder encodes the residual bit-stream using either a plain Huffman coder or an adaptive one. Similarly, the expander module consists of three sub-modules; namely, a shifted Huffman decoder, a bit-stream-to-integer converter, and a reconstructor, the purposes of which are self-explanatory.

There are two program sub-modules that are shared by all the coders; namely, Huff.c and bitio.c, which are used for Huffman coding/decoding and input/output of bit-streams, respectively. The remaining modules are:

sadpcm-raw-c.c (SADPCM compression algorithm for raw images),

sadpcm-raw-e.c (SADPCM expansion algorithm for raw images),

sadpcm-pbm-c.c (SADPCM compression algorithm for .PBM images),

sadpcm-raw-e.c (SADPCM expansion algorithm for .PBM images),

hbadpcm-raw-c.c (HBADPCM compression algorithm for raw images),

hbadpcm-raw-e.c (HBADPCM expansion algorithm for raw images).

The compiled programs are called sadpcm-raw-c, sadpcm-raw-e, sadpcm-pbm-c, sadpcm-pbm-e, hbadpcm-raw-c, and hbadpcm-raw-e, respectively. Each of the above programs can be executed as follows:

program-name input output,

which assumes "input" as the file-name for the original image and "output" as the file-name for the compressed image. However, the programs for raw images also prompt for the row and column dimensions of the input image.

3.6 Testing of software codes

The performances of the above programs were tested on a wide variety of test images, which are 8-bit deep, i.e., digitized to 256 gray levels. Some samples of the test results are provided in Tables 1 and 2. In our study, the compression ratio is measured by:

$$\text{Compression Ratio} = \frac{\text{Original file size}}{\text{Compressed file size}}.$$

The sample of five test images chosen for SADPCM consists of: two radiographs, called R1 and R2, respectively, and three others are standard test images available at a number of internet sites. Most of the standard pictures in our experiments are taken from the test-image database at the RPI site. These are referred to as Lena, Pepper, and Flower, respectively. All the pictures selected for this experiment, with the exception of Flower, are (512x512) in size. The picture, Flower, is of size (480x512).

Table 1 shows the lossless compression results using SADPCM-RAW-C and a block size of 32x32. As can be readily seen, the SADPCM method provides good compression for a wide variety of

test images.

For HBADPCM, the block size is chosen to be 32x32, and both a_1 and a_2 are estimated using the RLS method. Table 2 provides a sample of the test results. The results clearly indicate that in terms of lossless compression efficiency, HBADPCM coders perform very well.

Finally, Figs. (3a) and (3b) demonstrate the effectiveness of the simple interpolation scheme for the missing blocks, presented in Section 3.4.3 above. Fig. 3a depicts the missing blocks of Pepper in Level 1, assuming a block size of 8x8, whereas Fig. 3b shows the interpolated picture. As is apparent from these figures, even a simple interpolation scheme produces reasonably good quality of interpolated picture. Of course, block sizes have to chosen small for achieving good results.

4. IMPROVEMENT OF SADPCM AND HBADPCM CODING PERFORMANCE USING CONTEXTUAL SOURCE CODERS

Researchers have shown that there are essentially two ways to achieve good lossless compression; namely, 1) using a very efficient image model to decorrelate the data and then encoding the residuals by a simple, memoryless entropy coder, such as, first-order Huffman coder [21], or 2) using a simple model to partially decorrelate the image data and then employing a highly efficient contextual source coder to encode the residuals. A good example of approach 1 is SADPCM [27], which is a suboptimal version of the 2-D SMAR coder, introduced in [34]. On the other hand, examples of the second approach include the powerful contextual source coders presented in [29]-[31].

There are advantages and disadvantages shared by both of the above approaches. When approach 1 is used, one can employ a very large, spatial prediction mask to remove most of the linear dependencies among the neighboring pixels, but the subsequent employment of a simple entropy coder can only capture the memoryless, i.e., first-order, entropy of the residual signal source. When approach 2 is used, although not much linear correlation gets removed by the simple image modeler, most of the nonlinear and high order probabilistic interdependencies can be removed by using a sophisticated contextual source coder. The main problem with the latter approach, however, is that it increases the complexity of the coder significantly.

From the above discussion, it is clear that an ideal strategy to achieving good lossless compression would be to marry a good, parsimonious image modeler with a reasonably good contextual source coder. That is exactly what we pursue here. When such an approach is used, there is hope of removing most of the linear dependencies by using an efficient, spatial prediction mask, which will decorrelate the data to the extent that a relatively low order contextual source coder will be sufficient to remove the remaining probabilistic interdependencies.

In this study, we chose the I-FADPCM coder, described in Section 3.3 above, as the predictive image modeler. As far as the contextual source coder is concerned, after conducting some initial experiments involving higher order arithmetic coders (HOAC) [31], contextual arithmetic coders (CAC) [29],[30], higher order Huffman coders (HOFC), and contextual Huffman coders (CHC) [35], we decided to use the CHC as our source coder because of three reasons:

i) higher order coders require too much computation;

ii) CHC turned out to be easier to work with, because all of our other programs are based on Huffman coders; and

iii) usually, the performance of CHC seems to be quite close to that of CAC.

The CHC introduced below is based on a novel coding strategy that seems to provide very good performance.

4.1 Description of the Contextual Huffman Coder (CHC)

The contextual Huffman coder (CHC) proposed here is based on the following idea. Suppose we have divided the image into smaller (MxM) blocks, and found the prediction residuals for each block. If we could generate and transmit a Huffman codebook for each individual block, then significant compression gain will result, because the block-by-block coders are now tuned to the local probability distributions of the residuals. This concept is illustrated in Table 3, which shows the compression gain that results from a gradual reduction of the block size, M.

The main difficulty in implementing the above idea is that the extra bits needed to transmit the block-by-block codebooks eats up any gain that results from the block-size reduction. To solve this problem, a novel CHC, as described below, is proposed to be used here.

For coding the residuals of any block, say, the (m,n)th one, first define a context window, which forms the search horizon for constructing the probability distribution of (m,n)th block residuals. In this study, we restrict our context window to be the nearest four causal neighboring blocks of the (k,l)th one, i.e., the blocks numbered (m,n-1), (m-1,n-1), (m-1,n), and (m-1,n+1). Next, the residual symbols of the block (m,n) are divided into two categories; namely, i) symbols that are common to both the (m,n)th block and its context window, referred to hereafter as the matching symbols (MAS) and ii) the ones that belong to the (m,n)th block, but not to its context window, referred to hereafter as the missing symbols (MIS). The MAS and MIS are coded using two separate codebooks, called as *Codebook "MAS"* and *Codebook "MIS"*, respectively. Whereas the *Codebook "MAS"* varies from block to block, the *Codebook "MIS"* is common for all of them. The construction of these codebooks is described briefly in the following subsections.

4.1.1 *Construction of Codebook "MIS"*

This codebook is basically a shifted Huffman coder based on the entire ensemble of the missing symbols gathered from all the blocks, and therefore, an initial pass is necessary to construct it. The transmitter transmits this codebook to the receiver as a side information. In our study, we used a shifted Huffman coder which is designed for the integer symbols pertaining to a fundamental range of [-126,126] and two special symbols known as the shift-up and shift-down, respectively.

4.1.2 *Construction of Codebook "MAS"*

The codebook "MAS" (for the matching symbols) of the (m,n)th block is basically a shifted Huffman coder based on the probability distribution of the context window's symbols (PDCWS) and the probability of occurrence of a missing symbol (POMIS). Notice that the POMIS requirement stems from the need to generate a prefix codeword to tell the receiver that the codeword following it pertains to a missing symbol. Whereas PDCWS can be generated on-the-fly by both the transmitter and the receiver, POMIS may be either transmitted as a block-by-block side information or approximated using the POMIS of the neighboring blocks. In this study, we follow the latter route and approximate the POMIS of the (m,n)th block as the mean of the POMIS values of its four nearest causal neighbors, because this avoids the necessity of transmitting any side information.

4.2 Experimental Results

The above SADPCM-CHC algorithm was tested on a wide variety of images. A sample of the experimental results obtained with three images, called Lena, Pepper, and Building, is shown in Table 4. As the results indicate, the proposed coder performs significantly better than SADPCM, albeit with some increased computational complexity. Our current research efforts are geared toward reduction of the overall computational burden.

5. EFFICIENT CODING SCHEMES FOR THUMBNAIL IMAGES

The thumbnail images, used currently at Rome Laboratory, are obtained by simply subsampling their originals by a factor of about four to eight. Although such images are quite small in size, their collective volume is often large enough to pose significant problems for efficient transmission, and therefore, further compression of such images becomes highly desirable. However, the compression of the currently used thumbnail images is made difficult by the facts that: i) there is very little correlation (or, redundancy) left among the neighboring pixels of such images, and ii) compression of such images using simple methods often give rise to coding artifacts, such as, granular noise and false contours. Therefore, improved techniques for both generation as well as compression of thumbnail images are deemed to be

quite important, and we explore some of them here.

This study focusses on three different approaches to achieve either higher compression or better quality of thumbnail images; namely, i) using simple, but more efficient, compression schemes, ii) employing simple image enhancement filters, and iii) using improved subsampling strategies for generation of such images. The results presented below seem to indicate that good compression gain can indeed be realized by employing a judicious combination all the above approaches.

A new, simple and efficient image compression technique, which seems to be well suited for compression of thumbnail images, is presented first. Comparative performance results with respect to its better known (but, more complicated) counterpart, JPEG, are also given. And, then issues related to compression of thumbnail images are addressed.

5.1 A Simple Image Model for Lossy Compression

For compression of a large volume of thumbnail images, one requires a simple and efficient compression strategy to start with. In this regard, two alternative strategies, which immediately stake their claims, are: i) differential pulse code modulation (DPCM) and ii) JPEG compression algorithm. Although DPCM is very easy to implement, its performance is usually significantly poorer than JPEG's. Thus, over the past few years, JPEG has gradually emerged to be the de facto standard for lossy image compression.

In spite of the above scenario, however, considerable amount of research is still being devoted to: i) improve the quality of JPEG-compressed images, and ii) devise new techniques that promise to offer better performance versus complexity trade-offs than JPEG. In this Section, we present one such algorithm which attempts to achieve the last of the two above objectives.

Our new lossy image compression algorithm is actually a significantly improved version of the classical DPCM method [1],[2], but retains the computational simplicity of the latter. The key to its improved performance lies in a simple and novel technique that is employed to adapt the coefficients of a conventional DPCM coder from pixel to pixel. The new algorithm is simply called adaptive DPCM (ADPCM) and it is based on a so called spatially varying DPCM (SVDPCM) image model, as discussed below.

5.1.1 SVDPCM Image models and parameter estimation

Typically, a 2-D digitized image is regarded as a nonstationary 2-D signal consisting of pixel intensity values, $\{f(i,j), 1 \le i \le L, 1 \le j \le L\}$, where i denotes the row index and j stands for the column index. A fixed 2-D DPCM image model, fitting this data, has the form [5]:

$$f(i,j) - 0.95\,[f(i,j-1) + f(i-1,j)] + 0.9025\, f(i-1,j-1) = w(i,j) \qquad (17)$$

where w(i,j) is the modelling error, which is regarded as a zero-mean, 2-D white noise sequence.

As compared to above, a SVDPCM image model can be regarded as a spatially varying version

of (17), which has the form:

$$f(i,j) + a_1(i,j)f(i,j-1) + a_2(i,j)f(i-1,j) + a_1(i,j)a_2(i,j)f(i-1,j-1) = w(i,j), \quad (18)$$

where $a_1(i,j)$ and $a_2(i,j)$ denote the spatially varying model coefficients. Whereas equation (17) allows straightforward prediction of f(i,j) in terms of its three neighboring pixels, doing the same from (18) requires estimation of $a_1(i,j)$ and $a_2(i,j)$ for all i and j. Although a variety of techniques can be used to achieve this, in order to keep the computational complexity of the algorithm to a bare minimum, we use a simple steepest decent strategy that minimizes the cost criterion,

$$J = |e(i,j)| = |f(i,j) - \hat{f}(i,j)| \quad (19a)$$

where |e(i,j)| denotes the absolute value of the error between f(i,j) and its predicted value based on f(i,j-1), f(i-1,j) and f(i-1,j-1), which is given by

$$\hat{f}(i,j) = -a_1(i,j)f(i,j-1) - a_2(i,j)f(i-1,j) - a_1(i,j)a_2(i,j)f(i-1,j-1). \quad (19b)$$

Notice that |e(i,j)| can be written as,

$$|e(i,j)| = \mathrm{sgn}[e(i,j)]e(i,j),$$

where sgn(x) equals +1 for x>0 and -1 for x<0. Therefore, differentiating |e(i,j)| with respect to $a_1(i,j)$ and $a_2(i,j)$, and setting them to zero, we obtain:

$$\delta e(i,j)/\delta a_1(i,j) = -\delta\hat{f}(i,j)/\delta a_1(i,j) = 0, \quad (20a)$$

$$\delta e(i,j)/\delta a_2(i,j) = -\delta\hat{f}(i,j)/\delta a_2(i,j) = 0, \quad (20b)$$

which, in view of (19b), yield

$$-f(i,j-1) - a_2 f(i-1,j-1) = 0, \quad (21a)$$

$$-f(i-1,j) - a_1 f(i-1,j-1) = 0. \quad (21b)$$

Thus the optimal estimates of $a_1(i,j)$ and $a_2(i,j)$ are given by:

$$\hat{a}_1(i,j) = -f(i-1,j)/f(i-1,j-1), \quad (22a)$$

$$\hat{a}_2(i,j) = -f(i,j-1)/f(i-1,j-1). \quad (22b)$$

The above estimates can be utilized to derive a predictive image coding scheme as follows.

5.1.2 Lossy image coding based on SVDPCM image model

The predictive coding scheme essentially consists of a one step ahead predictor of y(k), an adaptive quantizer for the residual errors, and an encoder. The one step ahead prediction involves calculation of $\hat{a}_1(i,j)$ and $\hat{a}_2(i,j)$ from (6), assuring stability of the predictive quantizer [12], and constructing a predicted value of f(i,j) from (19b).

Notice that equation (19b) can also be rewritten in the form of a 2-D multiplicative autoregressive (MAR) model [12],

$$a(q_1,q_2)\, f(i,j) = w(i,j), \quad (23a)$$

where q_1^{-1} and q_2^{-1} denote the unit backward shift operators along the columns and the rows, respectively

and

$$a(q_1,q_2) = (1 + a_1(i,j)q_1^{-1})(1 + a_2(i,j)q_2^{-1}). \tag{23b}$$

Therefore, assuming $\hat{a}(q_1,q_2)$ denote the estimated model, the predictive coder can be succinctly described by [12]:

$$p(q_1,q_2) = 1 - \hat{a}(q_1,q_2), \tag{24a}$$

$$\tilde{f}(i,j) = p(q_1,q_2)\hat{f}(i,j), \tag{24b}$$

$$d(i,j) = f(i,j) - \tilde{f}(i,j), \tag{24c}$$

$$\hat{d}(i,j) = \text{quantize}[d(i,j)], \tag{24d}$$

$$\hat{f}(i,j) = \tilde{f}(i,j) + \hat{d}(i,j), \tag{24e}$$

where $p(q_1,q_2)$ is called the predictor, $\tilde{f}(i,j)$ is the predicted value of $f(i,j)$, $\hat{f}(i,j)$ is the coded pixel value, $d(i,j)$ denotes the prediction error, and $\hat{d}(i,j)$ is the quantized prediction error. For the purpose of simplicity, a three-level block adaptive quantizer [12],[38] is used. In this case, the quantized prediction error is given by:

$$\hat{d}(i,j) = \begin{cases} \Delta, & d(i,j) > \alpha, \\ 0, & -\alpha \le d(i,j) \le \alpha, \\ -\Delta, & d(i,j) < -\alpha, \end{cases} \tag{25a}$$

where

$$\Delta = D\sigma_e \text{ and } \alpha = B\sigma_e, \tag{25b}$$

and σ_e^2 denote the prediction error variance computed over an MxM block. The parameter D controls the dynamic range of the quantizer and the tradeoff between granular noise and peak-clipping, whereas the parameter B determines the entropy of the quantized prediction error signal, $\hat{d}(i,j)$ [12],[38]. Increasing the value of B reduces the entropy of $\hat{d}(i,j)$ but causes the signal-to-noise-ratio to deteriorate. Finally, the quantized residuals are coded using a fixed-to-variable-length block code (FVBC) [12],[38].

Next, we address the issue of stability of the predictive quantizer. The stability criteria of 2-D MAR predictive coders have been analyzed in detail by Burgett and Das [12], the results of which are directly applicable here. In particular, we have the following stability result.

Lemma 1

i) The predictive quantizer given by (24a)-(24e) is stable provided the 2-D MAR model, (23), is BIBO stable, and ii) the BIBO stability of (23) is assured provided $|a_1(i,j)|<\lambda$ and $|a_2(i,j)|<\lambda$, $0<\lambda<1$, for all i and j.

Proof

The proof is given in [39]. For purpose of brevity, it is omitted here.

Next, we develop a robust predictive coding scheme by utilizing the above lemma. First, notice

that the substitution of (22a) and (22b) into (19b) yields a predicted value of f(i,j) as,

$$\tilde{f}(i,j) = f(i,j-1)f(i-1,j)/f(i-1,j-1). \tag{26}$$

In light of Lemma 1, however, the above predicted value has to be corrected if the criteria stated in Lemma 1 are not met. The corrected predicted value is computed by setting

$$a_k(i,j) = -\lambda, \text{ if } |a_k(i,j)| \geq \lambda, \; k = 1, 2,$$

and correcting the computation of $\tilde{f}(i,j)$ as necessary. As is easily shown, the procedure for computation of the corrected $\tilde{f}(i,j)$ can be summarized as follows:

- First, compute the estimates of $\hat{a}_1(i,j)$ and $\hat{a}_2(i,j)$ from (22a) and (22b).
- If both $|a_1(i,j)| \geq \lambda$ and $|a_2(i,j)| \geq \lambda$, compute $\tilde{f}(i,j)$ as

$$\tilde{f}(i,j) = \lambda[f(i,j-1) + f(i-1,j)] - \lambda^2 f(i-1,j-1), \tag{27}$$

  else compute it from (26).

The subsequent steps of the coding scheme are same as what is depicted in equations (24c)-(24e) and discussed earlier.

Next, some experimental results are presented.

5.1.3 Experimental results

To evaluate the performance of the image coding scheme described in the previous Section, three different images were selected. These are called Lena, Pepper and Jet. Each of these images is digitized to 256 gray levels, and constitutes of 512x512 pixels. Three quantitative performance measures were used; namely, mean square reconstruction error (MSRE), peak signal-to-noise-ratio (PSNR = $10\log_{10}(255^2/\text{MSRE})$), and average bit-rate (BR) [12]. For computation of the average bit-rate (BR), a FVBC scheme, as described in [12],[38], was used. The input block size for FVBC was chosen to be 4 and variable-length Huffman codes, as discussed in [12],[38], were used to encode the residuals.

The parameters, D and B, of the three-level center-clipping quantizer were varied from one experiment to another. The value of the first set of experiments uses the Lena image with 16x16 block size and D fixed at 2.0. The value of B was then varied from 1.3 to 1.7. In this case, Table 5 shows the variation of SNR and BR. As expected, the bit-rate decreases, but the granular noise degrades the picture quality as B is increased. In the next experiment, D and B were kept fixed at 2.0 and 1.3, respectively, and the block size is varied. Table 6 summarizes the variation of MSRE, SNR and BR with block size. In general, larger block sizes decrease the effectiveness of the block adaptive quantizer which results in lower reconstructed picture quality. Finally, the performance of ADPCM was tested on two additional images, namely, Pepper and Jet, using 16x16 block size with D = 2.0 and B = 1.5. Table 7 summarizes the values of MSRE, SNR and BR for different pictures.

The performance of the algorithm was also compared with that of JPEG. These results are shown

in Table 8. Notice that the performance of ADPCM is slightly inferior compared to that of JPEG. However, it must also be borne in mind that JPEG's computational complexity is higher than that of ADPCM. As an example, for an MxM image block, the computation of DCT (needed by JPEG) using a fast algorithm would require approximately $4M^2\log_2 M$ multiplications, whereas ADPCM would need about $5M^2$ multiplications.

5.2 Enhancement Filters for Image Compression

As far as image compression is concerned, low-pass filters are often found to be useful for: i) generating subsampled images, ii) enhancement of coarsely quantized images, and iii) preprocessing an image to remove unnecessary details before coding. All of these aspects are important for compression of thumbnail images, because such images often result from decimation, coding, and quantization of the original pictures.

With the ultimate objective of simplicity in mind, this investigation focusses on only two kinds of low-pass filters; namely, i) zero-phase smoothing or averaging filters, and ii) median filters. Both of them use either 3x3 or 5x5 sliding windows centered over a pixel. The smoothing filter replaces each pixel value by its local mean computed over the window, whereas the median filter replaces it by the local median. Since false contours often result from coding and coarse quantization of an image and the median is robust in presence of them, the median filters often enhance such images better than the mean filters.

5.3 Subsampling Strategies

The choice of a good subsampling strategy is very important for efficient generation and compression of thumbnail images, because it has a direct bearing on both quality and compressibility of such images. In this study, we made a comparative evaluation of three alternative subsampling strategies; namely, i) plain down-sampling, ii) mean-pyramid down-sampling, and iii) median-pyramid down-sampling.

The plain down-sampling simply involves subsampling by a factor of say, 4 to 6, along both rows and columns. The mean-pyramid down-sampling, on the other hand, consists of successive stages that involve simple averaging followed by subsampling. The averaging can be accomplished using a variety of sliding masks or windows. For example, one of the popular schemes uses a nonsymmetric 2x2 window to compute the average of {f(m,n), f(m,n+1), f(m+1,n), f(m+1,n+1)} before performing subsampling. However, in this study, we have used a 3x3 symmetric window that computes the average value of f(m,n) and its eight nearest neighbors before performing subsampling. It should also be mentioned that the mean-pyramid down-sampling scheme is basically a special case of Laplacian pyramid down-sampling [8], where each successive stage consists of a low-pass filter followed by a down-sampler.

Finally, the median-pyramid down-sampling scheme is basically a variant of the above technique,

where the mean is replaced by the median computed over a sliding window. Although usually both mean and median pyramid schemes perform equally well, the median one may work better if the high resolution image contains granular noise, false contours, or other artifacts that often result from coding and coarse quantization.

5.4 Choice of an Efficient Scheme for Thumbnail Image Compression

As mentioned before, the primary goal of this study is to find compression strategies for thumbnail images that either offer better performance versus complexity trade-off, or improve quality of the reconstructed images. In order to achieve these goals, several alternative strategies were investigated. Essentially, these involve judicious combinations of the strategies outlined in Sections (5.2)-(5.4) above and these can be described as,

i) plain subsampling and compression (PSC) scheme (which is currently used at Rome Laboratory);

ii) plain subsampling, compression, and interpolation (PSCI) scheme;

iii) mean-pyramid subsampling and compression (MPSC) scheme;

iv) median-pyramid subsampling and compression (MEDPSC) scheme;

v) compression and plain subsampling (CPS) scheme;

vi) compression and mean-pyramid subsampling (CMPS) scheme;

vii) compression and median-pyramid subsampling (CMEDPS) scheme.

In the PSC scheme, the original image is first subsampled (by a factor of say, 4) using the plain down-sampling strategy and then the subsampled image coded using either JPEG/ADPCM, or some other compression technique. The PSCI is a slight variation of the above one, which attempts to gain higher compression by first down-sampling to a level lower than the required one (say, by a factor of 8 when the required reduction factor is only 4), compressing the lowest resolution image and subsequently interpolating it to obtain the required thumbnail image. The MPSC scheme is similar to PSC except for the replacement of plain-subsampling by mean-pyramid subsampling. The MEDPSC scheme a simple variant of MPSC, where the mean-pyramid is replaced by the median-pyramid one. Finally, the CPS, CMPS and CMEDPS are variants of PSC, MPSC, and MEDPSC, respectively, where the order of performing subsampling and compression are switched.

The performance of the above strategies were evaluated using both quantitative and qualitative methods. For quantitative performance evaluation, we used the measures of PSNR and BR, as described earlier, whereas for qualitative evaluation, we used the measure of perceived quality as judged by independent observers. The main results can be summarized as follows.

- among the seven schemes, the PSCI scheme performs worst;

- the performance of MPSC and MEDPSC are quite similar to each other and both allow higher compression than PSC, with very little loss of subjective quality;
- CPS, CMPS, and CMEDPS deliver higher compression than their alternative schemes, (where the order of compression and subsampling is switched), i.e., PSC, MPSC, and MEDPSC, respectively;
- CMEDPS is found to be the best scheme among the methods considered.

Finally, examples of thumbnail images constructed using plain subsampling, PSC and CMEDPS, are provided in Figures (4a)-(4d). Fig. (4a) shows the original 512x512 Pepper image, whereas Fig. (4b) is the thumbnail image obtained after plain subsampling by a factor of 4, which gives a compression of 16. Figures (4c) and (4d) show the thumbnail images obtained using PSC and CMEDPS, respectively, both of which result in a compression ratio of about 45. As a comparison of PSC and CMEDPS clearly indicates, CMEDPS gives an image of much better quality than PSC.

6. Conclusion

This project focuses on three main objectives; namely, i) development of usable software codes for two new image compression algorithms called SADPCM and HBADPCM, ii) further improvement of their compression performance by incorporating context-dependent source coders, and iii) development of an efficient scheme for coding and transmission of thumbnail images, used at Rome Laboratory and elsewhere within the Air Force. All of the above goals have been realized. The specific achievements can be summarized as follows:

a. Usable "C" codes have been developed and tested for SADPCM and HBADPCM algorithms;

b. In order to improve performance of the above coders, a new contextual Huffman coding algorithm has been developed and tested;

c. A new, efficient lossy image compression technique has been developed;

d. Several strategies for compression of thumbnail images were investigated and a novel scheme called CMEDPS has been shown to perform better than the rest.

It is hoped that the results of this investigation will be useful to Rome Laboratory and other Air Force organizations.

Acknowledgement

The author gratefully acknowledges the support and help provide by the following organizations and individuals:

- AFOSR and Oakland University for providing financial support;
- Mr. J. Nethercott and F. Rahrig for general guidance and help;
- Mr. R. Nallapati for programming and testing a majority of the algorithms;
- Mr. J. Anand for fruitful discussion and programming/testing of SADPCM-CHC.

References

[1] A. K. Jain, "Image Data Compression: A Review," Proceedings of the IEEE, Vol. 69, pp. 349-389, March 1981.

[2] M. Rabbani, editor, *Selected Papers on Image Coding and Compression*, SPIE Optical Engineering Press, Bellingham, Washington, 1992.

[3] W. B. Pennebaker, *JPEG Still Image Data Compression Standard*, Van Nostrand Reinhold, New York, 1993.

[4] R. B. Arps and T. K. Truong, "Comparison of International Standards for Lossless Still Image Compression," Proc. IEEE, Vol. 82, No. 6, pp. 889-899, June 1994.

[5] A. K. Jain, *Fundamentals of Digital Image Processing*, Prentice Hall, Englewood Cliffs, NJ, 1989.

[6] A. Gersho and R. M. Gray, " Image Coding Using Vector Quantization," Proc. IEEE International Conference on Acoustics Speech and Signal Processing, pp. 428-431, April 1982.

[7] H. M. Hang and J. W. Woods, "Predictive Vector Quantization of Images," IEEE Transactions on Communications, Vol. COM-33, pp. 1208-1219, November 1985.

[8] P. J. Burt and E. H. Adelson, "The Laplacian Pyramid as a Compact Image Code," IEEE Transactions on Communications, Vol. COM-31, No. 4, pp. 532-540, April 1983.

[9] M. Todd and R. Wilson, "An Anisotropic Multi-Resolution Image Data Compression Algorithm," Proc. IEEE International Conference on Acoustics Speech and Signal Processing, pp. 1969-1972, April 1989.

[10] S. R. Burgett and M. Das, "Predictive Image Coding Using Multiresolution Multiplicative Autoregressive Models," IEE Proceedings-I, Vol. 140, No. 2, pp. 127-134, 1993.

[11] J. W. Woods and S. D. O'Neil, "Subband Coding of Images," IEEE Transactions on Acoustics, Speech and Signal Processing," Vol. ASSP-34, pp. 1278-1288, October 1986.

[12] S. R. Burgett and M. Das, "Predictive Image Coding Using Two-Dimensional Multiplicative Autoregressive Models," Signal Processing, Vol. 31, No. 2, 1993.

[13] P. Roos et al, "Reversible Intraframe Compression of Medical Images," IEEE Transactions on Medical Imaging, Vol. 7, No. 4, pp. 328-336, 1988.

[14] M. Das and S. Burgett, "Lossless Compression of Medical Images Using Two-Dimensional

Multiplicative Autoregressive Models," IEEE Trans. on Medical Imaging, Vol. 12, No. 4, pp. 7221-726, 1993.

[15] M. Das and C. C. Li, "Simple Space-Varying Least Squares Model for Lossless Medical Image Compression," Electronics Letters, Vol. 30, No. 11, pp. 849-850, 1994.

[16] N. Tavakoli, "Lossless Compression of Medical Images," Proceedings of Fourth Annual IEEE Symposium on Computer-Based Medical systems, pp. 201-207, 1991.

[17] V. K. Heer and H-E Reinfelder, "A Comparison of Reversible Methods for Data Compression," SPIE Vol. 1233, Medical Imaging IV: Image Processing, pp. 354-365, 1990.

[18] H. Blume and A. Fand, "Reversible and Irreversible Image Data Compression Using S-Transform and Lempel-Ziv Coding," SPIE Vol. 1091, Medical Imaging III: Image Capture and Display, pp. 2-17, 1989.

[19] L. Wang and M. Goldberg, "Comparative Performance of Pyramid Data Structures for Progressive Transmission of Medical Imagery," SPIE Vol. 1232, Medical Imaging IV: Image Capture and Display, pp. 403-413, 1990.

[20] L. Wang and M. Goldberg, "Reduced-Difference Pyramid: a Data Structure for Progressive Transmission," Optical Engineering, Vol. 28, No. 7, pp. 708-716, 1989.

[21] D. A. Huffman, "A Method for Construction of Minimum Redundancy Codes," Proc. IRE, Vol. 40, pp. 1098-1101, 1952.

[22] R. G. Galleger, "Variations on a Theme by Huffman," IEEE Transactions on Information Theory, Vol. 24, No.6, pp. 668-674, 1987.

[23] A. Lempel and J. Ziv, "Compression of Two-Dimensional Images," *Combinatorial Algorithms on Words*, pp. 1410-154, Springer-Verlag, 1985.

[24] I. H. Witten, R. M. Neal, and J. G. Cleary, "Arithmetic Coding for Data Compression," Communications of the ACM, Vol. 30, No. 6, pp. 520-540, June 1987.

[25] M. Das and D. L. Neuhoff, "Near-Lossless Compression of Digitized Images," Communications and Signal Processing Lab Report #283, University of Michigan at Ann Arbor, June 1993.

[26] M. Das, D. L. Neuhoff, and C. L. Lin, "Near-Lossless Compression of Medical Images," Proc. IEEE International Conf. on Acoustics Speech and Signal Processing, held in Detroit, Michigan, 1995.

[27] M. Das, F. W. Rahrig, and J. Nethercott, "An Efficient Suboptimal Adaptive DPCM Scheme for Lossless and Near-Lossless Image Compression," Oakland University Technical Report No. TR-95-ESE-08-01, revised, October 1995.

[28] M. Das, J. Nethercott, and F. W. Rahrig, "A Suboptimal Block-Adaptive DPCM Scheme for Hierarchical Lossless Compression of Medical and Other Images," Oakland University Technical Report No. TR-95-ESE-08-02, revised, October 1995.

[29] P. Howard and J. Vitter, "New Methods for Lossless Image Compression Using Arithmetic Coding," Information Proc. and Management, Vol. 28, No. 5, pp. 765-779, 1992.

[30] P. Howard and J. Vitter, "Error Modeling for Hierarchical Lossless Image Compression," Proceedings of 1992 Data Compression Conference, J. Storer and M. Cohen, eds., pp. 269-278.

[31] T. V. Ramabadran and K. Chen, "The Use of Contextual Information in the Reversible Compression of Medical Images," IEEE Trans. on Medical Imaging, Vol. 11, No. 2, June 1992, pp. 185-195.

[32] S. E. Elnahas, K. Tzou, et al, "Progressive Coding and Transmission of Digital Diagnostic Pictures," IEEE Transactions on Medical Imaging, Vol. MI-5, No. 2, pp. 73-83, 1986.

[33] K. Tzou, "Progressive Image Transmission: a Review and Comparison of Techniques," Optical Engineering, Vol. 26, No. 7, pp. 581-589, 1987.

[34] M. Das and S. Burgett, "Reversible Compression of Medical Images Using Space-Varying Multiplicative Autoregressive Models," Proc. 1993 European Conf. on Circuit Theory and Design, held in Davos, Switzerland, during August 30 - September 3, 1993.

[35] H. C. Huang and J. L. Wu, "Windowed Huffman Coding Algorithm with Size Adaptation," IRE Proceedings-I, Vol. 140, No. 2, pp. 109-113, April 1993.

[36] A. Narula and J. S. Lim, "Error Concealment Techniques for an All-Digital High-Definition Television System," SPIE, Vol. 2094, pp. 304-315, 1993.

[37] K-H Jung, J-H Chung, and C. W. Lee, "Error Concealment Technique Using Projection Data for Block-Based Image Coding," SPIE, Vol. 2308, pp. 1466-1476, 1994.

[38] B. S. Atal, "Predictive coding of speech at low bit rates", IEEE Trans. Comm., 1982, Vol. 30, pp. 600-614.

[39] M. Das and C. L. Lin, "A Simple Adaptive DPCM Technique For Predictive Image Compression", Oakland University Technical Report No. TR-96-ESE-06-01, June 1996.

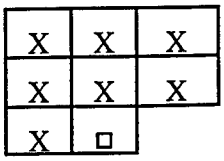


Figure 1. Causal window for estimation of $r_{fl}(1,0)$ and $r_{fl}(0,0)$ for each pixel. □ denotes current pixel, X denotes window pixels.

4	0	2	0	4	0	2	0	4
0	1	0	1	0	1	0	1	0
2	0	3	0	2	0	3	0	2
0	1	0	1	0	1	0	1	0
4	0	2	0	4	0	2	0	4
0	1	0	1	0	1	0	1	0
2	0	3	0	2	0	3	0	2
0	1	0	1	0	1	0	1	0
4	0	2	0	4	0	2	0	4

Figure 2. Hierarchical ordering of blocks. Blocks "4" are coded first, followed by blocks "3", and so on.

Images	Compression Ratio using SADPCM-RAW-C
R1	3.23
R2	3.28
Lena	1.66
Pepper	1.64
Flower	2.35

Table 1. Compression ratios achieved using SADPCM-RAW-C

Images	Compression Ratio using HBADPCM-RAW-C
R1	3.01
R2	3.13
Lena	1.64
Pepper	1.62
Flower	2.41

Table 2. Compression ratios achieved using HBADPCM-RAW-C

Image	Block Size	Compression Ratio
Lena	32x32	1.75
	16x16	1.81
	8x8	1.97

Table 3. Compression gain achievable by reducing the block size

Images	Block Size	Compression Ratio Using SADPCM-CHC	Compression Ratio Using SADPCM
Lena	32x32	1.73	1.66
	16x16	1.83	1.68
	8x8	1.97	1.69
Pepper	32x32	1.70	1.65
	16x16	1.82	1.68
	8x8	2.12	1.70
Bldg	45x45	1.62	1.60
	30x30	1.64	1.61

Table 4. Compression gain achievable using SADPCM-CHC

Values of B	MSRE	PSNR	BR
1.3	45.24	31.58	1.23
1.5	56.47	30.61	1.01
1.7	68.71	29.76	0.85

Table 5. Variation of MSRE, SNR and BR with B (D=2.0 in all cases); Image: Lena

BLOCK	MSR	PSNR	BR
64X64	56.97	30.57	1.20
32X32	50.40	31.11	1.21
16X16	45.24	31.58	1.24

Table 6. Variation of MSRE, SNR and BR with block size (fixed D=2.0, B=1.3); Lena.

IMAGE	MSRE	PSNR	BR
PEPPER	69.72	29.70	0.83
LENA	56.47	30.61	1.01
JET	30.00	33.36	1.07

Table 7. Performance of ADPCM across several images (16x16 blocks; fixed D=2.0,B=1.5)

IMAGE	ADPCM (D=2.0, B=1.5)			JPEG		
	MSRE	PSNR	BR	MSRE	PSNR	BR
PEPPER	69.72	29.70	0.83	62.35	30.18	0.86
LENA	56.47	30.61	1.01	41.32	31.97	0.92

Table 8. Comparative Performance of ADPCM and JPEG

Figure 3a. Missing blocks of Pepper in Level 1

Figure 3b. Interpolated Pepper from available blocks of Level 1

Figure 4a. Original Pepper image

Figure 4b. Thumbnail of Pepper using plain subsampling
(Compression ratio = 16)

Figure 4b. Thumbnail of Pepper using plain subsampling and compression (PSC)
(Compression ratio = 45)

Figure 4c. Thumbnail of Pepper using compression and median pyramid
subsampling (CMEDPS) (Compression Ratio of about 45)

MODE-LOCKED FIBER LASERS

Principal Investigator Joseph W. Haus
Department of Physics
Rensselaer Polytechnic Institute
Troy, NY 12180-3590

Final Report for:
Summer Research Extension Program
Rome Laboratory

Sponsored by
Air Force Office of Scientific Research
Bolling, AFB, Washington, DC
and
Rome Laboratory

February 16, 1997

Mode-locked Fiber Lasers, Joseph W. Haus

Abstract

The grant was used to accomplish the following tasks:

Technical accomplishments

Computations and analysis of mode-locked fiber lasers were performed. This work was begun during the summer research programs and continued during the year. Our model includes periodic amplification and loss, which is necessary for simulating side-band radiation observed in experiments. Our studies showed that the side-bands observed in the pulse spectrum could be controlled by using a technique we dubbed dispersion balancing, which manages the dispersion in each segment of the laser cavity. We also made a thorough analysis of a related Cr^{4+}:YAG laser; the results were quantitatively compared with experiments. Our fiber laser simulations are the first with propagation in throughout the laser; it provided us with a tool to observe and to quantitatively examine the side-bands appearing on the pulse spectra.

Three papers were prepared under the auspices of the summer program and have been accepted for publication in refereed journals. In addition the results were presented at three conferences.

Business Report

1. Regular trips to the Photonics Laboratory were made to discuss progress toward developing experimental mode-locked fiber lasers. An experiment was planned to synchronize two fiber lasers and lock their phases together. Other experiments were planned including a cross-correlation experiment for the next phase of the experiments. Experimental work delayed until the pump laser with higher power and a better modal emission pattern arrived.

2. Numerical calculations were performed by James Theimer at Rome Laboratory to compare against analytical calculations done at Rensselaer Polytechnic Institute on multiple-scales averaging of equations for pulse propagation in birefringent fibers. The vector soliton solution compared well between the two methods. However, several analytical solutions generated by a Lie group method were found to be numerically unstable. Nevertheless, the numerical solutions were of interest for further study and James Theimer has been examining them further.

3. Recent work on a Cr^{4+}:YAG mode-locked laser at the Photonics Laboratory by Dr. Mark Krol came to my attention. A model was constructed using the master equation approach. James Theimer wrote a program to solve the master equation and I took the task of determining the analytical properties of the model. We posit that the laser is dominated by soliton shaping and its pulse width is largely determined by properties of the saturable absorber. This model gave good agreement with the experiments. It provides a useful benchmark for further modeling of mode-locked fiber lasers, which are in operation at the Photonics Center of Rome Labs.

1 Introduction

Erbium-doped fiber lasers have many cavity designs and operation regimes, which are being explored as a source of high repetition rate, energetic, ultrashort pulses, that meets the needs of future communication networks. Ultrashort pulses are generated using a technique called mode-locking[1]. Passive mode-locking, using a fast saturable absorber-like action, has produced pulses of sub-picosecond duration; the pulses are soliton-like, i.e. hyperbolic secant shaped; this is significant because solitons have proven to be robust against the presence of losses and amplification in fiber transmission systems; i.e. they are stable against the presence of small perturbations. The topic of soliton transmission in optical fibers has rapidly evolved from a pure research topic to an emerging technology through a series of important technological breakthroughs (overcoming challenging obstacles) in long-distance, high bit-rate communication systems. In addition, soliton interactions have been proposed for logic and routing devices, which can perform important information processing tasks [2].

Three years ago we began a project devoted to numerical simulations to study pulse propagation in optical fibers with special emphasis on mode-locked fiber laser operation. Our goal was to provide accurate modeling to help resolve issues regarding the stability of the pulse propagation. Our analysis included several laser designs: the figure-eight fiber laser cavity, whose overall length is large compared to nonlinear shaping mechanisms in the cavity, a Cr^{4+}:YAG solid-state laser designed with a saturable absorber mirror, and a straight fiber-optic cavity with a saturable absorber mirror.

A short-pulse laser requires saturable absorber-like action to assist in achieving shorter pulses. The saturable absorber transmits more of the higher intensity portion of the pulse while the lower intensity experiences higher losses. There are now several fiber-optic photonic devices that can be applied to achieve the fast saturable absorber-like action and there is already an extensive literature developed about erbium-doped fiber amplifier [3] and erbium-doped fiber laser [4]. The optimum configuration of the devices in each laser design requires a detailed analysis.

2 Figure-eight Fiber Laser

Figure–eight lasers are of special interest due to their ability to produce nearly transform limited pulses at a very high repetition rate. Their operation was first demonstrated in 1990 [5, 6] with two mode–locking configurations. One used the nonlinear optical loop mirror (NOLM)[5] and the other uses a nonlinear amplifying loop mirror (NALM) by Richardson[7] and by Duling[6]. Bulushev[8] simulated the NOLM based

laser by direct integration of the nonlinear Schrödinger equation (NSE) in the NOLM, but no propagation in the amplifier, which was modeled as a homogeneously-broadened, saturable gain medium. Tzelpis et al.[9] reported a similar analysis on the NALM based figure–eight lasers. They can also be analyzed as Additive-pulse mode-locking (APM) lasers[10].

It has been experimentally found that periodic perturbations to a soliton-like pulse produce spectral sidebands [11, 12, 13, 14, 15]. The research of Dennis et al.[15] has quantitatively examined this phenomenon. The side bands are a result of dispersive wave shedding that circulates in the cavity and is amplified along with the pulse when the wavelength is phase matched. The operation of the figure–eight laser is sensitive to the total dispersion in the cavity, which includes the amplifier section of the laser.

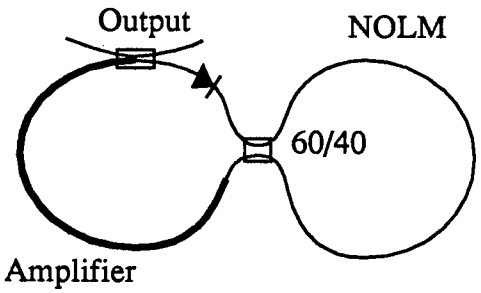


Figure 1: Sketch of the figure–eight laser geometry showing the important elements. The isolator in the amplifier section is depicted as a diode symbol. The NOLM section has counter-propagating pulses.

2.1 Figure-eight Laser Model and Simulation

For our research project [16], we initially modeled the behavior of a fiber laser mode-locked by an nonlinear optical loop mirror (NOLM). Pulse propagation was simulated by means of the split-step Fourier transform method[17]. The amplifier was simulated as a two-level atom with a parabolic gain curve. The population inversion was taken to be a constant and gain saturation was neglected. A schematic of the laser is shown in Figure 1. The loop mirror has a length of four soliton periods. A 60/40 directional coupler was used in the center, and 10% of the pulse energy was coupled out of the cavity. We found that the pulse experienced a loss which was dependent on the length

of the amplifier. We could not obtain a stable pulse if the length of the amplifier was longer than 0.83 dispersion lengths.

We attributed this instability to the competing tendencies of the amplifier and the NOLM. The two pulses in the NOLM favor a fundamental soliton shape; maximum transmission is produced for a pulse with a field envelope with a shape 2.2sech(τ), where the intensity and τ are in soliton units. For this situation the pulse will be transmitted with only a small amount of distortion. The two pulses at the output of the NOLM are recombined and are not a fundamental soliton in the amplifier. When the fiber amplifier is extended, the pulse is reshaped toward a fundamental soliton and for a long amplifier, it approaches the fundamental soliton shape. We attributed the length dependent loss to the fact that this reshaping process resulted in a pulse shape which was not correct for complete transmission through the NOLM.

The fiber amplifier was modeled as a two-level system with a parabolic line shape as in Ref.[18]. We consider the case of a 60/40 directional coupler between the two cavity sections in Figure 1 and a 10 % output coupler. Initially, the NOLM loop is four soliton periods long in scaled units. Duling[6] found that this length produces minimum loss and pulse distortion. As the pulse width shortens during successive round trips in the cavity, the soliton period correspondingly shortens. It was assumed that the population inversion in the amplifier was uniform, and for steady–state conditions it totally recovers between passes of the pulse through the amplifier. The equation describing pulse propagation through the amplifier is

$$i\frac{\partial E}{\partial z}+\frac{1}{2}\frac{\partial^2 E}{\partial \tau^2}+E|E|^2=i\frac{G}{2}E+i\mu\frac{\partial^2 E}{\partial \tau^2}, \tag{1}$$

The equations have been scaled to soliton units throughout [17]. The time is scaled to a value T_0, related to the initial pulse width and the field is scaled to a value E_0; the length is scaled by the dispersion length $L_D = T_0^2/|\beta_2|$, where β_2 is the group velocity dispersion parameter; it is negative in the wavelength regime near 1.5 μm; $\beta_2 \approx -20$ ps^2/km. An often used length related to L_D is the *soliton* period whose definition is $Z_0 = \pi L_D/2$. The field amplitude scaling corresponds to a fundamental soliton , $\delta|E_0|^2 = |\beta_2|/T_0^2$, where δ is related to the fiber's Kerr nonlinearity and the fiber's effective core area. The gain parameter G is a variable in our simulations and the gain dispersion parameter $\mu = GT_2^2$ is a product of the gain parameter and the polarization relaxation time we used is $T_2 = 100fs$ in physical units, which is appropriate for erbium–fibers. In numerical simulations all our scaled parameters are based on scaling $T_0 = 300fs$.

The left hand side of Eq. (1) has the elements of dispersion and nonlinearity required

for pulse propagation in an optical fiber. Our pulse lengths are long enough that effects, such higher-order dispersion, stimulated Raman scattering and self-steepening, are negligible. This portion of the evolution equation is applied to propagation of the two pulses in the NOLM.

The two additional terms on the right hand side of Eq. (1) are included in the simulation when the pulses propagate through the amplifier section of the laser. They describe a gain curve with a maximum at the pulse's center frequency and a parabolic gain profile; the gain profile is much wider than our pulse spectra and this approximation is not a limiting factor.

The saturation energy of the amplifier can be given by $E_s = h\nu_0 a_{eff}/2\sigma$. This expression can be found, for example, in reference [18]. In this expression ν_0 is the central frequency of the pulse, a_{eff} is the effective core area, and σ is the absorption cross section of the laser line. The parameters which effect saturation energy are, therefore either not under our control, or are severely constrained by amplifier design criteria. Typically, saturation energy will be found to be around 10 mJ. This implies that the values of saturation energy used in references [8] and [9] were far too large to be physically realistic.

First we investigated the effect of the fiber amplifier's length. Altering the length changes the amount of dispersion in the cavity, which also limits the minimum pulse length in the cavity. We found that stable pulses could not be produced if the amplifier was longer than 0.7-0.8 dispersion lengths; we attribute this effect to reshaping the pulse in the laser amplifier, since, during this portion of the evolution, the pulse has a higher energy than a fundamental soliton energy and the pulse tends to shed energy while it evolves toward a fundamental soliton (this will be discussed further below). Since the NOLM loop is 4 soliton periods long, this implies a total loop length of about 7.0 dispersion lengths. For amplifiers shorter than 0.8 dispersion lengths, we could only achieve stable, single pulse operation for a narrow band of values for the gain coefficient. If the gain was either too great, or too small, the pulse would decay away. These findings are summarized Figure 2.

We also found a maximum amplification, beyond which a single pulse in the loop will not be stable. For higher gains the cavity would adjust to include two pulses in the cavity. The effect of a second pulse in the loop will be to cut the gain coefficient in half, which would reduce the gain to the point where stable pulses could again be formed. This implies a pump dependent transition from a regime with one stable pulse in the loop, to one with multiple pulses[19, 20]. Such a transition is observed in NALM based lasers.

The directional coupler for the NOLM loop produces two pulses of roughly half the

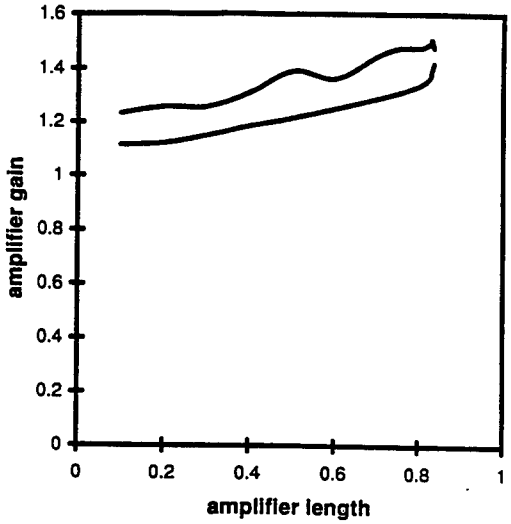


Figure 2: Stable operation regime for a single pulse in the figure–eight laser.

input intensity, traveling in opposite directions. This would make them approximately unit solitons, and as such subject to minimum distortion during propagation. However, when these pulses are recombined and propagate in the amplifier, they tend to be reformed into a soliton shape suitable to their amplitude, and to shed some light to dispersive wave radiation. The degree of transmission in the NOLM loop is highly dependent on pulse shape, and the reshaped pulse will experience greater loss than the original sech-shaped pulse. If this loss becomes too great, the laser will simply cease to function, as we found in our simulation.

The spectrum of this pulse is shown in Figure 3. The central section of the spectrum is hyperbolic–secant shaped, but there is additional structure in the wings that is due to dispersive–wave shedding. The shoulders are consistent with the placement of the sidebands given by the formula[12, 13]

$$\omega_m = \sqrt{8mZ_0/L - 1}; \tag{2}$$

where L is the length of the fiber and Z_0 is the soliton period corresponding to the steady–state pulse width.

The final pulse widths depend on several factors including, the amplifier gain and length, and the length of the NOLM. The higher gain parameter has a shorter width due to stronger nonlinear shaping mechanisms. We can also work toward shorter pulses by adjusting the cavity length. Naturally, as the pulses shorten other propagation loss effects[17], such as Raman scattering and higher-order dispersion, must incorporated

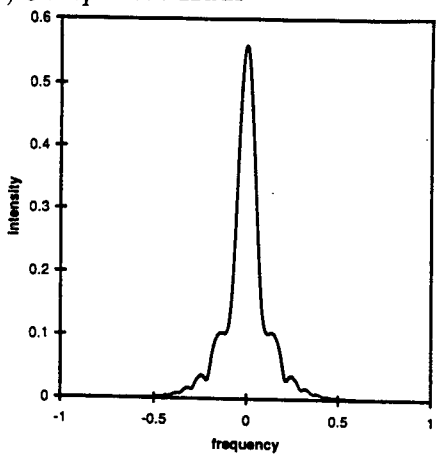


Figure 3: Spectrum of the pulse near the stability limit; the amplifier length is 0.7 in dispersion lengths and the gain is near the minimum of the stability region of Figure 2. The spectrum displays shoulders on the envelope that is attributable to dispersive–wave shedding.

into the analysis; and the amplifier model should incorporate a more accurate gain profile shape and the dynamical evolution of the active medium.

The cavity length of our F8L in steady state is around 4 soliton periods or greater, depending on the amplifier length and gain. This is somewhat larger than the F8L designed with a nonlinear amplifying loop mirror; in this geometry the loop mirror is shorter and the unidirectional cavity section is also shorter. They typically operate in a regime from 0.6 to 4 soliton periods. We coincidentally find that the maximum amplifier length operation corresponds to a cavity length of about 8 soliton periods, based on the pulse energy in each section of the cavity (see the discussion after Eq. (2)); at this point the pulse undergoes strong reshaping during propagation and it sheds a large fraction of its energy to the dispersive wave, which is strongly resonant with the solitary pulse and leads to its eventual break-up.

2.2 Dispersion Balanced Figure-eight Laser

Our analysis and understanding of the length dependent loss for the F8L suggests a possible alternate solution[21]. We should design the amplifier fiber so that the pulse is nearly a fundamental soliton in that fiber, as well as in the NOLM; this balances the two sections of the laser in Figure 1 so that the pulses are always close to a fundamental soliton shape and that perturbations of that shape are kept to a minimum. For a fundamental soliton the relation between the average power and the

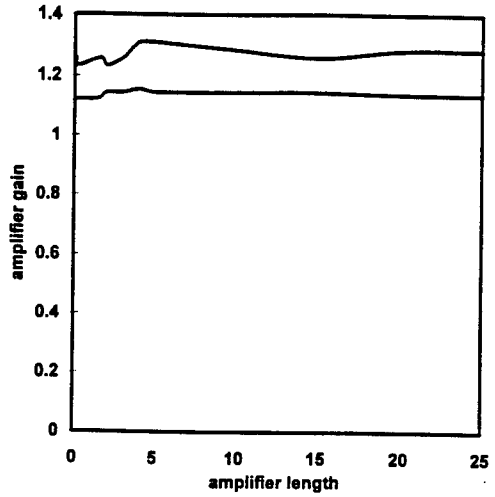


Figure 4: Stability regime for stable single pulse operation is a band in the amplifier gain between about 1.1 and 1.3.

pulse width parameter is fixed and given by

$$1 = \frac{\delta P_0 T_0^2}{|\beta_2|}, \tag{3}$$

where the parameters δ and T_0, were defined under Eq. (1) above, and $P_0 = E_0^2$ is the peak power. Since, we desire that T_0 be the same in both sections, then the dispersion must be increased by a factor of around 2.2 to compensate for the increased peak intensity at the output of the NOLM, when counter-propagating solitary waves combine in the amplifier section; see Figure 1.

This solution was incorporated into our laser model. Figure 4 shows the regions in which stable pulses could be formed. It should be noted that in this case the limit of the abscissa of the graph extends to 25 dispersion lengths; the pulse is very well approximated by a fundamental soliton and for further amplifier lengths, no change is expected. This is in stark contrast with Figure 2, where the cavity sections are not dispersion balanced. It should be further noted that the gain required for stable operation is quite constant, representing minimal gain dependent loss and a result of the soliton pulse shape. When the amplifier is longer than about 5 dispersion lengths, pulse reshaping is minimal, and the maximum and minimum stable gains are nearly constant. If the amplifier is shorter than this, the pulses deviate from a soliton shape and the output from the NOLM is sensitive to its input pulse shape.

To obtain further insight into the functioning of this laser, the pulse shape was

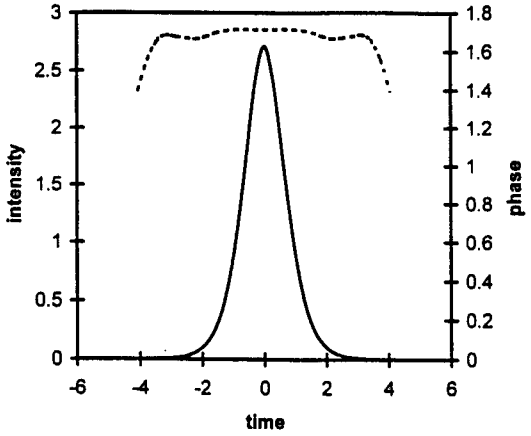


Figure 5: Pulse shape and the phase of the pulse after propagating through a 25 dispersion length amplifier. The minimum amplifier gain of 1.1 was used and the pulse is seen after the output coupler.

examined at the point at which it left the output coupler, and before it entered the directional coupler; it is shown in Fig. 5. The case that was examined was for a 25 dispersion length amplifier, with a gain parameter, $G = 0.0056$; This corresponds to an amplifier gain of $e^{GL} = 1.15$. The output pulse has a nearly uniform phase. The pulse intensity corresponds almost exactly to that of a fundamental soliton, with a pulse intensity of about 2.7. In comparing the relationship of the pulse width to pulse height it must be remembered that the ratio has been altered by a factor of 2.2.

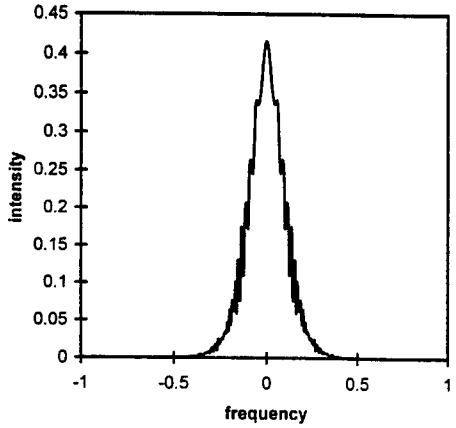


Figure 6: The pulse spectrum of the pulse in Fig. 5. The side bands are reduced in this balanced laser; their position is consistent with soliton perturbation theory, see Refs. [10].

Figure 6 shows the pulse spectrum. Sidebands have clearly developed. The separa-

tion of the first order sideband from the center is 0.05 in normalized frequency units, which is consistent with dispersive-wave shedding[12]. In the case of a F8L, the soliton period varies as the pulse travels through the laser. Since the length of the amplifier is 25 dispersion lengths, this section alone has a length far longer than eight soliton periods usually found to limit operation of the F8L; a restriction that also applied to our previous F8L simulation. This is significant as the length appears to represent a length limit for ring lasers and we attribute this to the improved rejection of the dispersive wave in our NOLM because the pulses are very nearly fundamental solitons. We also find that the dispersive wave component in the spectrum is increased when the cross-coupler splitting ratio is closer to 50/50 and the length of the NOLM is correspondingly increased.

The existence of a maximum amplifier length sets a limit of the shortest pulse that can be produced by a given laser. If the physical length of the amplifier is given by κ, and the maximum amplifier length in dispersion lengths is L_{max} the limit would be expressed as

$$\kappa = L_{max} L_D. \tag{4}$$

Since the dispersion length can be defined as $L_d = T_0^2/|\beta_2|$, Eq. (4) can be transformed to

$$\left(\frac{\kappa|\beta_2|}{L_{max}}\right)^{1/2} = T_0. \tag{5}$$

The limit on L_{max} is related to the dispersive wave soliton resonance of 8 soliton periods [15], but since this laser has reduced side band amplitudes, this limit no longer applies. Our results indicate, however, that it is possible to decrease the shortest possible pulse, while increasing the average dispersion in the laser, which is a possibility which has not been explored before. Previous work has centered on optimizing the performance of the laser by using fiber with normal dispersion [22]. It should be noted that our method for optimizing the laser could be implemented by decreasing the dispersion of the fiber in the NOLM loop.

3 Cr^{4+}:YAG Mode-locked Laser

Recently self-starting, passive mode-locking of a Cr^{4+}:YAG laser using a saturable absorber mirror (SAM) was reported[23, 24]. The SAM is designed from a quarter-wave stack of GaAs/AlAs layers with a double quantum well structure grown at the interface. This is an important element in the cavity design. Its absorption changes as the laser is tuned, which in turn affects the cavity mode-locking action. The successful

operation of the laser depends upon the specific features of this element. The prism pair in the cavity provides negative dispersion for solitonic pulse shaping in the cavity. The laser operating wavelengths are in the range from 1490 nm to 1540 nm, which lies in the transmission region for fiber optic communications. The output power varies between 40 and 80 mW, which is intense enough to launch pulses in fibers with sufficient energy to study multi-soliton propagation effects with ultra-short pulses.

Using a simple model we are able to accurately predict the mode-locking behavior of the Cr^{4+}:YAG laser. The cavity design is reduced to a small number of parameters that are independently measured or determined. This laser differs from the F8L designs in the previous sections since its total cavity length does not exceed a soliton period. This allows for a great simplification in the analysis, because averaged equations for the evolution of the field amplitude can be derived[1].

3.1 Cr^{4+}:YAG Model

The average soliton approach is viable to describe the properties of a Cr^{4+}:YAG laser. The equation in this regime is[25]

$$\frac{\partial a}{\partial z} = i(-D\frac{\partial^2 a}{\partial t^2} + \delta|a|^2 a) + D_3\frac{\partial^3 a}{\partial t^3} + (g-l)a + \frac{g}{{\Omega_f}^2}\frac{\partial^2 a}{\partial t^2} + \gamma_3|a|^2 a - \gamma_5|a|^4 a. \quad (6)$$

This equation is a form of the complex Ginzburg-Landau equation studied in several fields of physics and has terms besides those given in Eq. (1). The amplitude a represents the complex electric field envelope. D is the cavity dispersion, which is related to the second-order dispersion β_2 by $D = \frac{\beta_2 L}{2}$; L is the length of the medium that contributes to the dispersion; the third order dispersion coefficient is defined by $D_3 = \beta_3 L/6$. The Kerr nonlinearity, n_2, responsible for self-phase modulation, appears in the parameter $\delta = \frac{2\pi}{\lambda} n_2 P L_x / A_{eff}$; P is the peak power; A_{eff} is the effective area of the beam; λ is the wavelength; and L_x is the crystal length. The scaled value for the nonlinear parameter is $\delta = 0.8$. The cavity loss and gain is collected in the parameters l and g, respectively. The gain bandwidth is parameterized by Ω_f, whose value is $\Omega_f = 2.9 \cdot 10^{14} Hz$; and the saturable absorber is described by two parameters γ_3 and γ_5. The formation of bright solitons requires the second-order dispersion be negative; this is fulfilled by the dispersion compensating prisms. The first two parameters in parentheses are the soliton-shaping mechanisms of dispersion and nonlinearity. The detailed analysis is given in the manuscript[26].

The dominant pulse shaping mechanism in the cavity are due to dispersion and self-phase modulation terms. The form of the pulse amplitude is a hyperbolic-secant function

$$a = E_0 \mathrm{sech}(t/T_0) e^{iz/2L_D}. \tag{7}$$

The gain and saturable absorber terms determine the pulse energy, $W = 2E_0^2 T_0$, while these and higher-order dispersion terms perturb the pulse shape from the soliton solution. Important parameters are the average pulse width and the pulse energy, which are closely related by the dominant shaping terms:

$$T_0 = 4|D|/\delta W. \tag{8}$$

The pulse shape found by our simulations is close to a hyperbolic-secant form with deviations appearing in the wings. The energy is calculated from the area under the pulse. The balance between the energy in the linear and nonlinear gain-loss contributions to the energy is given by

$$l - g = \frac{2}{3}\gamma_3 E_0^2 - \frac{8}{15}\gamma_5 E_0^4 - \frac{1}{3}\frac{g}{\Omega_f{}^2 T_0^2}. \tag{9}$$

The last two terms, which are higher-order absorption saturation and gain dispersion terms, resp., make important corrections to this result and our results are consistent with this expression. This equation can be solved for the pulse width and the energy to determine consistency of our soliton shaping hypothesis represented by Eq. (8); this too is very closely followed and this consistency check is evaluated in the results. The deviations can be attributed to the third-order dispersion, which contributed to the continuum radiation.

The numerical computations were done using a split-step algorithm, which is described elsewhere and employed for our earlier studies of pulse propagation [16, 17, 27]. The linear coefficients were chosen as $l = 0.02$ and $g < l$ is adjusted to keep the steady-state pulse energy constant. Eq. (9) imposes restrictions on the maximum value of l; this is discussed later. In the experiments, the output coupler must be kept small to achieve mode-locking. Group delay dispersion and the third-order dispersion in the cavity are computed applying well tested techniques. The values for the cavity were calculated for the design in Ref. [24].

The quintic-order terms, γ_5, in Eq. (6) was required for stable operation. Without it the peak energy was unstable. This is also observed in Eq. (9), where the cubic term, γ_3, is balanced against the quintic term. The gain dispersion in the cavity is also important in the energy balance.

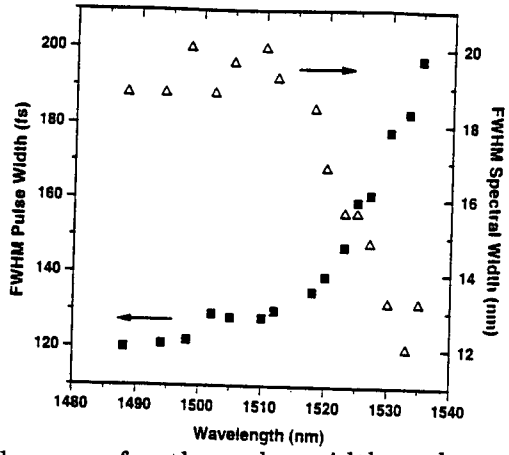


Figure 7: Experimental curves for the pulse width and spectral width versus wavelength. All values are full width at half maximum. After Ref. [24].

The experimental results for the pulse width and spectral width are reproduced from Ref.[24] and shown in Fig. 7. The trend in the data over the range of frequencies is mostly attributable to the variations in the dispersion, although there are also variations in the output energy. However, the curvature at longer wavelengths toward longer pulse widths exceeds expectations based on variations of the dispersion alone. This can be accounted for by the wavelength dependence of the absorption. The semiconductor material has absorption variations as the wavelength is tuned. In our case the change can be rather large, since the device operates close to the band edge of the semiconductor material. The roll off of the band-edge absorption strongly affects the ability of the laser to achieve pulse operation.

The simulation results in Fig. 8 display two situations. The solid curve exhibits the pulse width when the cavity dispersion is varied according to the values provided by calculations[26]; in addition, the solid line is the data for the saturable absorber parameter $\gamma_3 = 0.04$ and the long-dashed line is the data for $\gamma_3 = 0.02$. The results do not significantly change when third-order dispersion is omitted from the simulation. The cavity gain was adjusted at each frequency to keep the steady-state energy of the pulse constant. The pulse energy chosen was nearly the minimum for which stable pulses could be produced. this minimum is set by the requirement of satisfying Eq. (9). The pulse width is approximated from Eq. (8) and the results, which are represented by the short-dashed line in Fig. 8, are indistinguishable from the simulation points; in this case only the dispersion is varied. This is clearly shows the dominant soliton shaping mechanism for the model parameters. The pulse width value from the simulation is also in agreement with Eq. (9); deviations are within 6% and there are no fit parameters.

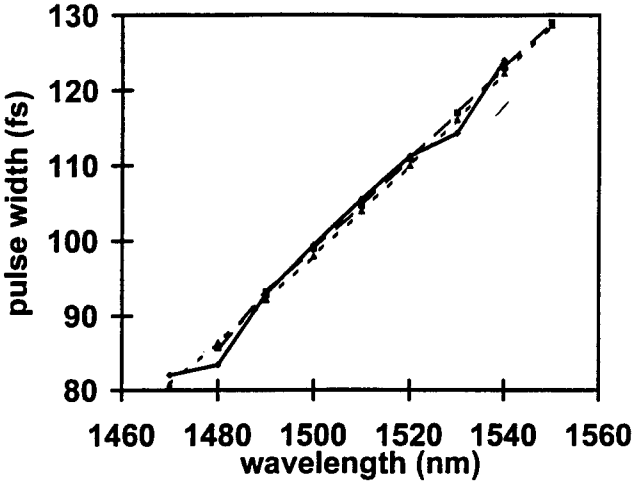


Figure 8: The pulse width versus wavelength from the simulations with change of the saturable absorber parameters.

We believe that the differences can be attributed to the third-order dispersion, which increases due to the cavity gain.

Simulations were refined to study the cause for the sharp upward trend of the pulse width as the laser was tuned to longer wavelengths, see Fig. 7. Simply maintaining a constant gain parameter, while letting the pulse energy vary, did not produce the curvature observed. Subsequently the sensitivity of the results to parameter variations was examined; the energy was held constant in these simulations by adjusting the gain. The loss was varied between 0.01 and 0.02; the gain bandwidth was changed between 2.9×10^{14} Hz and 1.93×10^{14} Hz. These changes were in addition to the saturable absorber values of 0.02 and 0.04 discussed for Fig. 8. The change in the gain needed to hold the energy constant did not steepen the trend for the pulse width versus wavelength; the pulse width was well described by Eq. (8). The mode-locking parameter, loss, and amplifier bandwidth are all poorly known for this system. They do not, however, appear to change the minimum pulse energy for stable operation. Since it is pulse energy that dictates the pulse width, and we know that the laser had barely enough gain to operate, these unknown parameters don't appear to affect our conclusions.

A portion of the excess change in the pulse width from the tuning curve of Fig. 7 is attributable to the change in the saturable absorber efficiency as a function of wavelength. This causes a variation of the output energy, which is consistent with the sharp upward turn of the pulse width at longer wavelengths.

When the absorption saturation variation is taken into account, the results are

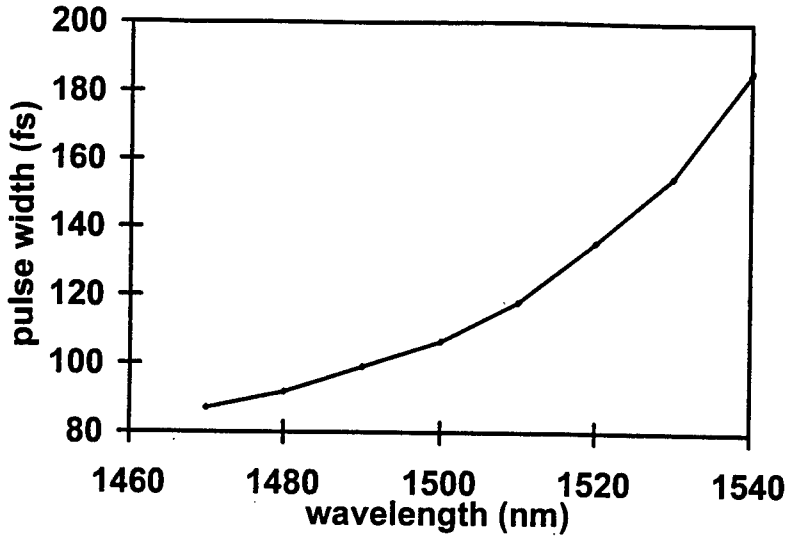


Figure 9: The pulse width versus wavelength from the simulations with change of the saturable absorber parameters.

modified. The curve of the pulse width versus wavelength in Fig. 9 demonstrates the effect that this has on the pulse width. The decrease in the saturable absorber parameters becomes severe at the long wavelengths, which results in a slow convergence to the steady-state solution. We were unable to find mode-locking for wavelengths longer than 1550 nm. The output energy cannot be held constant in these simulation by merely adjusting the gain parameter g.

The corresponding change of the output energy with wavelength is displayed in Fig. 10, where the minimum gain for achieving mode-locked pulses on the long wavelength side was used. The energy peaks near the band edge and drops by about 10 % on either side of 1520 nm. The saturable absorber change alone does not significantly change the pulse widths versus wavelength, but since it becomes less effective at longer wavelengths, it decreases the ability of the cavity to achieve mode locking operation.

The effect of gain saturation and the MQW band edge acting together are shown in Figs. 9 and 10, which show the pulse width and pulse energy as a function of wavelength. Since the effective gain is greater for less energetic pulses, stable pulses could be formed with smaller energies than in the previous cases. The pulse energy was set so as to have agreement between the pulse widths at 1550 nm had approximately the same width. Comparison of the result in Fig. 9 with that in Fig. 8, where gain saturation is neglected, shows a large contrast. The long wavelength pulses are much wider when gain saturation is included, since they can have smaller energies and still be stable, and pulse widths are relatively unchanged for short wavelengths. From this

we can conclude that gain saturation is an important factor in the excess pulse widths observed at long wavelengths. The energy in Fig. 10 is flat. Gain saturation helps to stabilize the output energy, but in the region where the saturable absorber becomes less effective, the pulse energy in Fig. 10 rolls of by nearly 20 %. This has a large affect on the pulse widths and is responsible for the excess increase at long wavelengths.

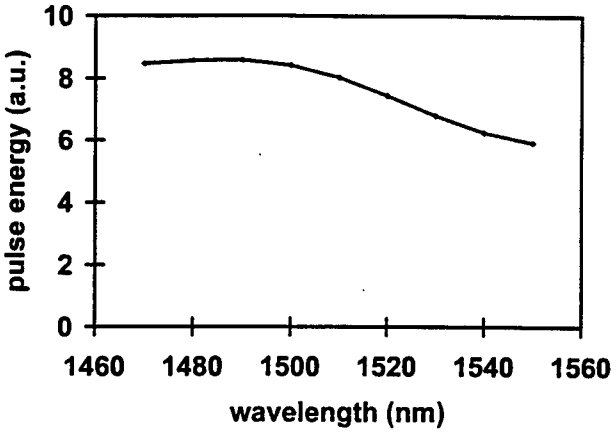


Figure 10: The energy of the steady-state pulses versus wavelength using the change of the absorption coefficient. The peak at 1520 nm occurs at the knee of the absorption curve, see Ref. [26].

We also found that the pulse spectra have exponential wings over six orders of magnitude[26]. The long wavelength spectrum has oscillations in the central region due to perturbations affecting the soliton propagation. The oscillations in the spectrum become stronger for longer wavelengths, where the third-order dispersion becomes stronger. Third-order dispersion causes the solitonic pulse to shed radiation to the continuum; turning off the third-order dispersion contribution decreases the oscillations found in the spectrum, although it does not eliminate them.

4 Conclusions

Erbium-doped Fiber lasers are compatible with present fiber communications technology at 1.5 μm. The laser can be an inexpensive source of ultrashort, intense optical pulses with a variety of optical communications applications for military C3I or commerce. Fiber lasers are being designed to be environmentally robust and inexpensive, which is a major impediment to wide scale applications. Lasers of this type, pumped by presently available laser diodes, could be inexpensive enough to be a pulse source in

every home and business; thus, it is possible that fiber to the home or office is not just a passive acceptor of high data rates, but also a source of high bit-rate information.

The research we began over the past three years was designed to help guide experimental development and to quantitatively determine the operating characteristics of fiber lasers. Our simulations will be used to optimize the design of laser parameters and try out different mode-locking elements in the laser cavity. The analytic tools we develop will help us to rapidly explore different operating regimes and help guide future experiments.

We made a detailed analysis of the figure-eight laser with a NOLM for a fast saturable absorber; its performance was optimized by balancing the dispersion in different parts of the cavity. This stabilized the laser output and made it less susceptible to dispersive-wave shedding and subsequent pulse beak-up.

In the future we will be able to assess the importance of material parameters, such as doping, third-order dispersion, stimulated Raman scattering, etc.. The threshold parameters can be examined and single pulse operation, pulse bursts and pulse trains, observed in experiments, can also be simulated. The research will provide a study of future fiber lasers that differ significantly from present ones, such as, the dispersion balanced figure-eight laser. The Cr^{4+}:YAG laser we analyzed and compared with experiments gave us further proof that the computational and analytical methods we developed are useful for designing and improving future mode-locked fiber lasers. This has led us to examine the operation of two new fiber lasers that were constructed at Rome Laboratories; the simulations of these new lasers is now underway and we expect that it will lead to improvements in design and better operation stability. We are also seeking to operate the lasers with low noise characteristics for future photonic applications in Air Force systems.

References

[1] I. N. Duling, ed., *Compact Sources of Ultrashort Pulses*, Cambridge University Press, Cambridge, (1995).

[2] M. Islam, *Ultrafast Fiber Switching Devices and Systems*, Cambridge University Press, Cambridge, (1992).

[3] E. Desurvire, *Erbium-doped Fiber Amplifiers: Principles and Applications*, Wiley, NY (1993).

[4] M. J. F. Digonnet ed., *Rare Earth Doped Fiber Lasers and Amplifiers*, Marcel Dekker, New York (1993).

[5] H. Avramopoulos et. al., "Passive modelocking of an erbium-doped fiber laser, Optical Amplifiers and Their Applications," Technical Digest Series., vol. 19, PDP 8, 1990.

[6] I. N. Duling, " All-fiber Modelocked Figure-eight Laser," OSA Annual Meeting 1990, PDP-4; "Subpicosecond All-fiber Erbium Laser," Electron Lett. **27**, 544 (1991).

[7] D. J. Richardson et. al., "Selfstarting, Passively Modelocked Erbium Fiber Ring Laser Based on the Amplifying Sagnac Switch," Electron. Lett. **27**, 542 (1991).

[8] A. G. Bulushev, E. M. Dianov, and O. G. Okhonikov, "Self-starting Mode-locked Laser with a Nonlinear Ring Resonator," Opt. Lett. **16**, 88 (1991).

[9] V. Tzelpis, S. Markatos, S. Kalpogiannis, Th. Schicopoulos, and C. Caroubalos, "Analysis of a Passively Mode-locked Self-starting All-fiber Soliton Laser," J. Lightwave Technol. **11**, 1729 (1993).

[10] H. A. Haus, E. P. Ippen and K. Tamura, "Additive-pulse Modelocking in Fiber Lasers," IEEE J. Quantum Electron. **30**, 200 (1994).

[11] N. Pandit, D. U. Noske, S. M. J. Kelly, and J. R. Taylor, "Characteristic Instability of Fibre Loop Soliton Lasers," Electron. Lett. **28**, 455 (1992).

[12] S. M. J. Kelly, "Characteristic Sideband Instability of Periodically Amplified Average Soliton," Electron. Lett. **28**, 806 (1992).

[13] N. J. Smith, K. J. Blow and I. Andonovic, "Sideband Generation Through Perturbations to the Average Soliton," J. Lightwave Technol. **10**, 1329 (1992).

[14] D. U. Noske, N. Pandit and J. R. Taylor, "Source of Spectral and Temporal Instability in Soliton Fiber Lasers," Opt. Lett. **17**, 1515 (1992).

[15] M. L. Dennis and I. N. Duling III, "Experimental Study of Sideband Generation in Femtosecond Fiber Lasers," IEEE J. Quantum. Electron., **30**, 1469 (1994).

[16] J. Theimer and J. W. Haus, "Figure-eight Fiber Laser Stable Operating Regimes," J. Mod. Optics , to appear (1996).

[17] G. P. Agrawal, "Nonlinear Fiber Optics", (Academic Press, NY, 1989).

[18] G. P. Agrawal, " Optical Pulse Propagation in Doped Fiber Amplifiers," Phys. Rev. A, **44**, 7493 (1991).

[19] D. J. Richardson, R. I Laming, D. N. Payne, V. J. Matsas, and M. W. Phillips, "Pulse Repetition Rates in Passive Self-starting, Femtosecond Soliton Fibre Laser," Electron. Lett. **27**, 1451 (1991).

[20] A. B. Grudinen, D. J. Richardson and D. N. Payne, "Energy Quantization in Figure Eight Fiber Laser," Electron. Lett. **28**, 67 (1992).

[21] J. Theimer and J. W. Haus, "Dispersion Balanced Figure-eight Fiber Laser," Opt. Commun. **134**, 161 (1997).

[22] K. Tamura, E. P. Ippen, H. A. Haus, and L. E. Nelson, "77-fs Pulse Generation from a Stretched-pulse Mode-locked All-fiber Laser," Opt. Lett. **18**, 1080 (1993).

[23] B. C. Collings, J. B. Stark, S. Tsuda, W. H. Knox, J. E. Cunningham, W. Y. Jan, R. Pathak and K. Bergman, "Saturable Bragg Reflector Self-starting Passive Mode Locking of a Cr^{4+}:YAG Laser Pumped with a Diode-pumped Nd:YVO_4 Laser," Opt. Lett. **21**, 1171 (1996).

[24] M. J. Hayduk, S. T. Johns, M. F. Krol, C. R. Pollock and R. P. Leavitt, "Self-starting Passively Mode-locked Femtosecond Cr^{4+}:YAG Laser Using a Saturable Absorber Mirror," preprint, (1996).

[25] H. A. Haus, "Short pulse generation", in I. N. Duling, III ed., Compact Sources of Ultrashort Pulses, (Cambridge Univ. Press, Cambridge, 1995), pp. 1-56. H. A. Haus, E. P. Ippen and K. Tamura, "Additive-pulse Modelocking in Fiber Lasers," IEEE J. Quant. Electron. **30**, 200 (1994).

[26] J. Theimer, J. W. Haus, M. Hayduk and M. Krol, "Mode-locked Cr^{4+}:YAG Laser: Model and Experiment," Opt. Commun., submitted (1996).

[27] J. W. Haus and J. Theimer, "Polarization Distortion in Birefringent Optical Fibers," Photonics Technology Letters **7**, 296 (1995).

John D. Norgard
Report not available at time of publication.

IMAGE MULTIRESOLUTION DECOMPOSITION AND PROGRESSIVE TRANSMISSION USING WAVELETS

Frank Y. Shih

Associate Professor

Department of Computer and Information Science

New Jersey Institute of Technology

Newark, NJ 07102

Final Report for:

Summer Faculty Research Program

Rome Laboratory

Sponsored by:
Air Force Office Of Scientific Research
Bolling Air Force Base, Washington, D. C.

December 1996

IMAGE MULTIRESOLUTION DECOMPOSITION AND PROGRESSIVE TRANSMISSION USING WAVELETS

Frank Y. Shih
Associate Professor
Department of Computer and Information Science
New Jersey Institute of Technology

Abstract

With the increased use of browsing remotely stored pictures over low bandwidth transmission lines, there is a need for better and better approaches. Because of the bandwidth constraints, various progressive image transmission techniques have been proposed. One such technique is computing Discrete Cosine Transform (DCT) of the image data and transmitting samples based on some priority determined by alternating current (AC) energy of transformed samples. In this technique, a rough version of the image will be given to the user and will be built gradually, providing an opportunity to terminate the transmission. The advent of wavelets reveals the power of providing multiresolution analysis. This report focuses on transmitting reduced resolution images, where the power of wavelets producing low resolution images is used. Next, the DCT for the reduced resolution image is calculated. The samples will be ordered for transmission according to the priority, which is determined by the AC energy of the transform domain samples. Our experiments indicate that the samples corresponding to lower resolution images get the highest priority for transmission, enabling quick recognition of the image to the user.

IMAGE MULTIRESOLUTION DECOMPOSITION AND PROGRESSIVE TRANSMISSION USING WAVELETS

Frank Y. Shih

1. Introduction

The browsing of remotely stored pictures over a low bandwidth transmission line results in poor performance particularly as the quantities of the image data get larger. If conventional line-by-line transmission is employed, it does not permit a rapid recognition of picture content, which would give the user the opportunity to begin to examine the picture or even interrupt the transmission at an early stage. The progressive transmission technique allows an approximate image to be built up quickly and the details to be transmitted progressively through several passes over the image.

The concept of progressive image transmission of digital pictures over low bandwidth transmission lines has some prospective applications, such as enabling a teleconference with interactive transmission where one wants to transmit visual material by first presenting a rough sketch with detail following. Another application would allow browsing through a picture database to find the picture of desired characteristics. For example, in an electronic radiology environment, a radiologist browses through many remotely stored pictures of a patient who has several diagnostic studies and several images per study. The efficient browsing would provide the ability to quickly abort the transmission of unwanted pictures as soon as they are recognized. Once the desired picture is identified, more information is transmitted until a clinical diagnosis is possible.

The introduction of wavelets in signal and image processing is a major breakthrough because of their ability to describe a signal or image in time and frequency simultaneously, thus overcoming classical limitations of Fourier analysis. If tracing back, some of the pioneering work of wavelet transform was introduced by Haar[1], Gabor[2], and Morlet [3]. As it came into the attention of other scientists it was recognized to be useful for other signal analysis applications. Nowadays, there exist widely-used wavelet transform algorithms (see [4], [5], [6]).

Like sines and cosines in Fourier analysis, wavelets are used as basis functions in representing functions. Given a time-varying signal $x(t)$, the wavelet series (WS) decomposes the signal into a basis of continuous-time wavelets as

$$x(t) = \sum_{j \in Z} \sum_{k \in Z} C_{j,k} \Psi_{j,k}(t) \qquad (1)$$

where $C_{j,k}$ denotes WS coefficients , j,k are integers, and $\psi_{j,k}(t)$ is called the 'analysis wavelet'. The WS coefficients $C_{j,k}$ are defined as

$$C_{j,k} = \int x(t)\, \Psi^*_{j,k}(t)\, dt \qquad (2)$$

where the asterisk stands for complex conjugate and $\psi_{j,k}(t)$ corresponding to scale 'a' and location 'b' is given by

$$\Psi_{a,b}(t) = 1/\sqrt{|a|}\, \Psi((t-b)/a) \qquad (3)$$

The time scale parameters (b,a) are sampled on a so-called "dyadic" grid in the time scale plane (b,a). A common definition is

$$a = 2^j,\ b = k2^j \qquad (4)$$

where k is an integer. The wavelets in this case are

$$\Psi_{j,k}(t) = 2^{-j/2}\, \Psi(2^{-j}t - k) \qquad (5)$$

The Discrete Wavelet Transform is very close to WS, but is applied to discrete time signals $x[n]$, where n is an integer, such that $n \in Z$. It achieves a multiresolution decomposition of $x[n]$ on J (is an integer) octaves, labeled by $j = 1, \ldots, J$, given by

$$x[n] = \sum_{j=1}^{\infty} \sum_{k \in Z} c_{j,k}\, h_j[n-2^j k] + \sum \sum b_{J,k}\, g_J[n-2^J k] \qquad (6)$$

The 'h' corresponds to synthesis wavelets, the discrete equivalents to Ψ. The 'h' stands for high pass. An additional lowpass term 'g' is used to ensure perfect reconstruction is called scaling sequences. The coefficients $c_{j,k}$ and $b_{J,k}$ are given by

$$c_{j,k} = \Sigma\, x[n] h^*_j[n-2^j k] \qquad (7)$$

$$b_{J,k} = \Sigma\, x[n] g^*_J[n-2^J k] \qquad (8)$$

Wavelet transform methods for image data produce a multiresolution representation, in which the spatial structure of the image is preserved in each level. Wavelets analyzes each dimension of the input data into two subbands: the first consists of the average information of the signal (low frequency) components and the second contains the detailed information (high frequency components). The low frequency information can be considered as an approximation of the image and the high frequency information as the details lost at this step of approximation process.

To illustrate the 2-D image wavelet transform by a separable 1-D decomposition algorithm, let $\varphi(x)$ be a 1-D low-pass filter and $\psi(x)$ be the mother wavelet function (a high-pass filter) associated with the scaling function. In two dimensions, the 2-D wavelet transform can be equivalently computed with a separable extension of the 1-D decomposition algorithm [7]. For a matrix of 2-D image data $f(x,y)$ of

dimension 2^n x 2^n , where n is an integer, we first convolve the rows of $f(x,y)$ with the 1-D wavelet filter and two matrices of dimension 2^n x 2^{n-1} are obtained as shown in Fig. 1, where $\varphi(x)$ and $\psi(x)$ are a low-pass filter and a high-pass filter, respectively.

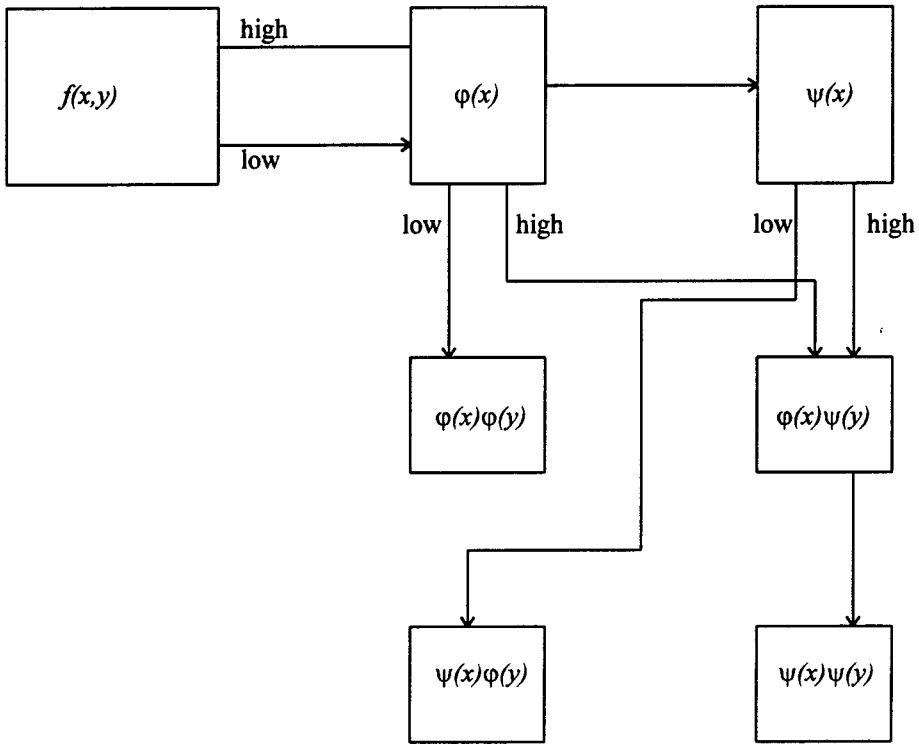


Fig. 1. The 2-D image wavelet transform by a separable 1-D decomposition algorithm.

We then convolve the columns of the resulting two matrices with the 1-D wavelet filter and the four resulting matrices $\varphi(x)\varphi(y)$, $\varphi(x)\psi(y)$, $\psi(x)\varphi(y)$, $\psi(x)\psi(y)$ of dimension 2^{n-1} x 2^{n-1} are obtained. The matrix $\varphi(x)\varphi(y)$ is a half-resolution down-sampled image of low frequencies (referred to as the thumbnail image), the image $\varphi(x)\psi(y)$ gives the vertical high frequencies (horizontal edges) and $\psi(x)\varphi(y)$ the horizontal high frequencies (vertical edges) and $\psi(x)\psi(y)$ the high frequencies in both directions (the corners as well as diagonal edges).

The filtering is repeated successively on the first, low-frequency subband, producing a pyramid of subbands. The output image of wavelet representation is shown in Fig. 2, where the top-left quadrant image is the output for the wavelet filter $\varphi(x)\varphi(y)$ (the low-frequency components), the top-right for $\varphi(x)\psi(y)$

(the horizontal edges), the bottom-left for $\psi(x)\varphi(y)$ (the vertical edges), and the bottom-right for $\psi(x)\psi(y)$ (the diagonal edges). Note that the subscript index indicates the level of resolution.

The three level representation for the lenna image obtained by applying Haar wavelet decomposition is shown in Fig. 3. We first apply the one-dimensional wavelet transform to each row of pixel values. This operation gives us average values along with the detail coefficients. Next, we treat these transformed rows as if they were themselves an image and apply the one-dimensional transform to each column. In Fig. 3, (a) represents the original image, (b) is the result of this row and column transformations. This is called level 1 decomposition. If repeat row and column transformations on the first quadrant of (b), we get the result as shown in (c), which corresponds to level 2 decomposition. If the above operations are repeated on the first quadrant of (c), we get the result as shown in (d), which corresponds to level 3 decomposition.

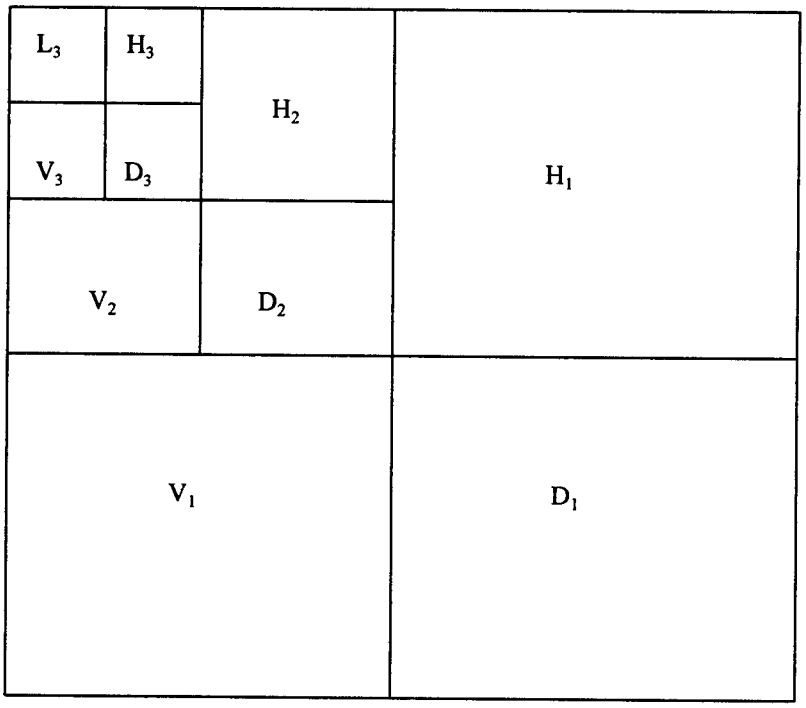
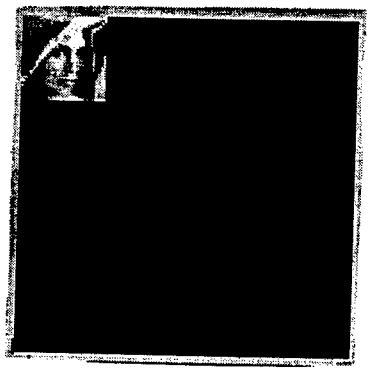


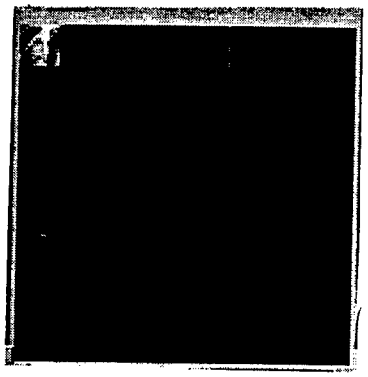
Fig. 2. Three level resolutions in a 2-D wavelet transform.

(a)

(b)



(c)



(d)

Fig. 3. (a) original image, (b), (c), (d) corresponds to level 1, level 2 and level 3 wavelet decompositions.

2. Methodology

This report deals with the application of ***Discrete Cosine Transform (DCT)*** [8] for the subbands, which are lower resolutions of the original image. All the transform samples will be transmitted in steps. This is achieved by developing classification and transmission step maps based on alternating current (AC) energy of transform domain samples. For the computation of the maps, the image is divided into sublocks of suitable size, then DCT of each subblock as well as AC energy is calculated. Classification map is obtained by grouping these subblocks based on the energy. Then in each group for the transmission of samples, the transmission step maps are used, which enables high priority samples, i.e., highest energy samples will be transmitted. This is called step wise transmission and the image is progressively reconstructed at the receiver.

Our main motivation for this approach is as wavelets generates subbands and recursive application of wavelets generates series of lower resolution images. The application of DCT and calculating AC energy reveals that the most of the energy is concentrated in the lower resolution images. Accordingly they will get highest priority for transmission resulting in a quick image reconstruction at the receiver. The classification maps of Fig. 3 (a), (b) are shown in Fig. 4 depicts this situation. The results are shown without applying wavelets (corresponding to Fig. 3 (a)) and with applying wavelets (corresponding to Fig. 3 (b)) and calculating DCT. For the purpose of calculation of DCT the image is divided into subblocks of a suitable size. For each subblock NxN the DCT is calculated as

$$F(u,v) = c(u)c(v) \sum_{j=0}^{N-1} \sum_{k=0}^{N-1} f(j,k)\cos(((2k+1)/2N)v\pi)\cos(((2j+1)/2N)u\pi) \qquad (9)$$

$$u = 0, 1, \ldots N\text{-}1, \quad v = 0,1, \ldots, N\text{-}1.$$

where $c(u) = 1/\sqrt{2}$, for $u = 0$,

$c(v) = 1/\sqrt{2}$, for $v = 0$,

$c(u) = 1$ for $u = 1, 2, \ldots, N\text{-}1$,

$c(v) = 1$ for $v = 1, 2, \ldots, N\text{-}1$,

j, k are integers,

$f(j,k)$ is the (j,k)-th value of the original subblock,

$F(u,v)$ is the DCT of the corresponding value.

3	3	2	1	1	0	2	3	3	2	1
3	2	1	1	1	0	3	3	0	1	0
2	1	2	1	0	3	3	3	0	0	3
1	2	2	0	2	2	3	3	0	0	3
2	2	0	1	0	0	2	0	1	0	3
1	2	0	2	1	2	1	1	1	0	2
0	0	0	3	3	3	1	1	1	0	0
0	0	1	3	3	3	1	1	2	0	0
0	2	2	3	3	2	1	0	3	1	2
0	3	3	2	3	2	2	1	2	1	3
1	3	3	2	2	3	1	0	2	1	2

(a)

1	0	0	1	0	0	2	1	2	2	1
0	0	0	1	0	0	1	1	3	2	1
1	0	0	0	0	0	1	1	2	1	1
0	0	1	0	0	0	1	3	2	2	1
0	0	1	0	0	0	2	3	2	1	1
0	0	0	0	0	0	3	2	2	2	2
2	1	1	3	1	2	2	2	2	2	2
2	1	1	2	2	3	2	2	3	3	3
1	1	3	3	3	2	1	3	3	3	3
1	3	3	1	3	1	2	3	3	3	3
2	3	3	2	3	2	3	3	3	3	3

(b)

Fig. 4. Classification maps. (a) is the result of directly computing DCT of the original image in Fig. 3 (a) and classifying into four groups. (b) is the corresponding result after applying level 1 wavelet decomposition. As can be seen from (b) activity is mainly concentrated in the first quadrant. Note : '0' represents the highest activity subblock.

3. Progressive Image Transmission

The block diagram of proposed progressive transmission system is shown in Fig. 5. In the progressive transmission of the transform domain samples, the transform domain samples with higher activity levels take larger information about image detail. In order to give higher priority for higher activity samples, as well as to keep the overhead in reasonable range, the subblock classification maps and transmission step maps are developed in determining the subsets of transform domain samples for stepwise transmission.

The design of subblock classification map and transmission step map is based upon three considerations. First, the direct current (DC) samples which represent the average brightness of subblocks will be transmitted at the first step. Second, the AC samples with higher magnitude having higher activity, will be transmitted with higher priority. Third, the order in which the transform domain samples are transmitted should be known by the receiving decoder such that the decoder can properly accumulate the progressively transmitted data to obtain the intermediate picture approximation. However, the ordering information is an overload for the transmission, thus should be kept within a reasonable range.

Classification Map: The activity levels of images are proportional to the transform domain samples AC energy. After performing the DCT on (m,l) subblock, the AC energy for the subblock is calculated as The calculated AC energy is accepted as the measure of the activity level of that subblock. After all the AC energy of subblocks are computed, they are sorted and classified into four groups according to this measure of activity. All sublocks in the same group are labeled by the same group number. This forms the classification map. The subblock classification map serve as an index to the transmission step maps and become a part of the overhead information needed by the receiver to decode the received data.

$$E(m,l) = \left(\sum_{u=0}^{N-1}\sum_{v=0}^{N-1} [F_{m,l}(u,v)]^2\right) - [F_{m,l}(0,0)]^2 \qquad (10)$$

where NxN is the subblock size, and (m,l) refers to the co-ordinate of the subblock.

Transmission Step Maps: The transmission step maps are used to store the transmission step level of all entries among four classification groups. One transmission step map will be set up per classification group. Each group contains subblocks of same energy and can be viewed as stacked to form a union block. An entry in a group refers to a sample from each stacked subblock. As discussed above, the DC samples of each subblock is the average subblock brightness, must be transmitted at the first step. We compute the activity level of each entry for all the groups except the DC entries. These entries will be sorted and four entries will be marked with the same step number 'i' to indicate that they will be transmitted at the i-th step. The transmission step maps are shown in Fig. 6.

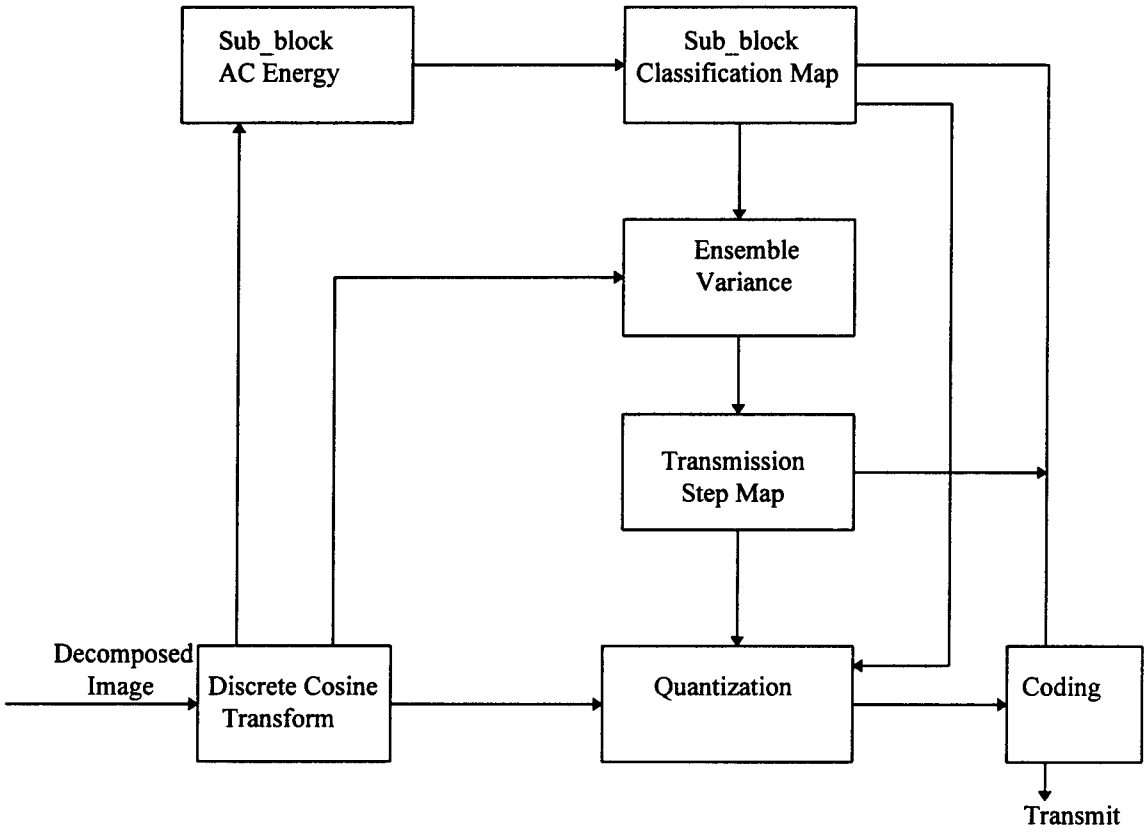


Fig. 5. A progressive transmission system.

Quantization and Coding : At each step appropriate samples will be selected and quantized, encoded and transmitted. A quantizer used at the transmitter to make the representation of transform coefficients to a shorter digital form fitted for transmission. The quantizer divides the whole input range into a finite number of quantization bands. For the input signals with amplitude lie in the quantization band , assigned with the same reconstruction level (set at the mid point of a band). The quantization error can be reduced by making the quantization step smaller. Whereas, the total number of quantization levels will grow larger and require more bits to encode. The distribution of coefficients at each step varies and thus an adaptive coding method is needed to encode the transformed coefficients.

TRANSMISSION STEP MAP 1 :

0	1	2	1	3	4	4	8	6	12	6
1	1	2	3	6	9	8	13	14	13	13
2	3	3	5	8	10	10	18	28	23	22
2	4	7	5	10	11	12	22	22	27	24
5	9	7	8	11	9	17	18	21	57	33
4	12	17	11	13	15	20	15	15	24	32
5	14	15	17	19	22	16	19	26	29	46
10	26	34	32	20	30	36	27	23	31	42
7	18	35	19	43	40	39	41	40	37	37
17	60	39	52	42	46	35	44	37	44	49
7	28	50	60	63	36	44	55	39	36	53

TRANSMISSION STEP MAP 2 :

0	6	9	18	16	21	31	37	33	23	26
11	12	16	29	32	20	24	34	42	33	41
14	14	19	16	24	29	25	65	21	42	38
28	26	20	27	21	43	41	44	48	64	62
60	30	61	40	33	31	28	25	50	62	65
62	45	43	47	38	29	25	51	30	45	46
38	56	62	48	58	59	61	47	32	30	23
45	35	67	50	49	54	45	49	43	38	40
70	54	51	61	68	63	47	36	68	48	25
72	70	66	59	89	63	58	52	34	41	47
74	82	81	67	57	57	52	46	48	35	27

TRANSMISSION STEP MAP 3 :

0	31	34	39	82	68	71	60	79	89	83
59	56	72	54	55	82	73	67	85	81	73
70	65	52	58	77	76	81	77	83	76	85
75	57	56	69	66	72	79	79	83	77	78
80	78	66	75	64	71	80	86	78	86	81
89	87	84	79	72	64	73	77	75	83	73
85	88	69	74	71	74	58	76	76	65	84
93	90	84	87	88	85	84	63	67	55	64
91	92	87	78	88	80	69	61	54	53	49
92	90	86	75	74	82	69	59	56	53	51
91	89	80	71	66	70	68	55	53	51	50

TRANSMISSION STEP MAP 4 :

0	91	110	99	115	118	118	106	111	117	120
86	91	104	94	94	110	96	100	114	101	120
109	92	98	106	116	119	120	103	117	105	97
92	88	96	118	102	109	105	113	113	119	115
95	113	105	108	105	115	114	102	116	99	112
108	99	98	103	120	101	108	106	102	116	111
111	93	112	100	112	99	111	100	112	104	107
95	94	110	115	109	97	109	101	94	98	102
87	90	103	119	101	95	107	117	103	97	108
96	93	119	96	107	113	114	117	104	104	107
90	95	100	98	110	106	114	118	116	97	93

Fig. 6 Transmission Step Maps of Fig. 3 (b) for the Four Classification groups.

At first, the two extremes of the transform coefficients are found and labeled as $MIN(k)$ for minimum, $MAX(k)$ for maximum. The input range is next computed as

$$R(k) = MAX(k) - MIN(k) \tag{11}$$

where 'k' denotes the transmission step. The number of bits required to encode the quantization output level is given by

$$B(k) = C \log_2 R(k) - D \tag{12}$$

where C and D are parameters which are adjusted in two considerations, one is the total number of bits desired to transmit and the other is the degree of quantization mean square error could be accepted. Once the $B(k)$ is determined, the number of quantization levels $N(k)$, the quantization interval $Q(k)$ and the smallest reconstruction level $Y0$ could be computed as follows.

$$N(k) = 2^{B(k)} \tag{13}$$

$$Q(k) = R(k)/N(k) \tag{14}$$

$$Y0(k) = MIN(k) + Q(k)/2 \tag{15}$$

The $Y0(k)$ and $R(k)$ will be transmitted through the channel ahead of the transform codes. They are the overload information for the receiver to de-quantize the transform codes at the kth intermediate image reconstruction.

4. Progressive Image Reconstruction

The image is progressively reconstructed at the receiver. A rough image is given first, then built up gradually. The transform codes received per step are decoded, de-quantized and inverse transformed. The block diagram of proposed reconstruction system is shown in Fig. 7. For each intermediate reconstruction step, the inverse descrete cosine transform is computed as

$$f(i,j) = (4/N^2) \sum_{u=0}^{N-1} \sum_{v=0}^{N-1} c(u)c(v)\, F(u,v)\cos(((2k+1)/2N)v\pi)\cos(((2j+1)/2N)u\pi) \tag{16}$$

$u = 0, 1, \ldots N\text{-}1, \quad v = 0, 1, \ldots, N\text{-}1.$

where $c(u) = 1/\sqrt{2}$, for $u = 0$,

$c(v) = 1/\sqrt{2}$, for $v = 0$,

$c(u) = 1$ for u = 1, 2, . . . , N-1,

$c(v) = 1$ for v = 1, 2, . . . , N-1,

j, k are integers,

$f(j,k)$ is the (j,k)-th value of the original subblock,

$F(u,v)$ is the DCT of the corresponding value.

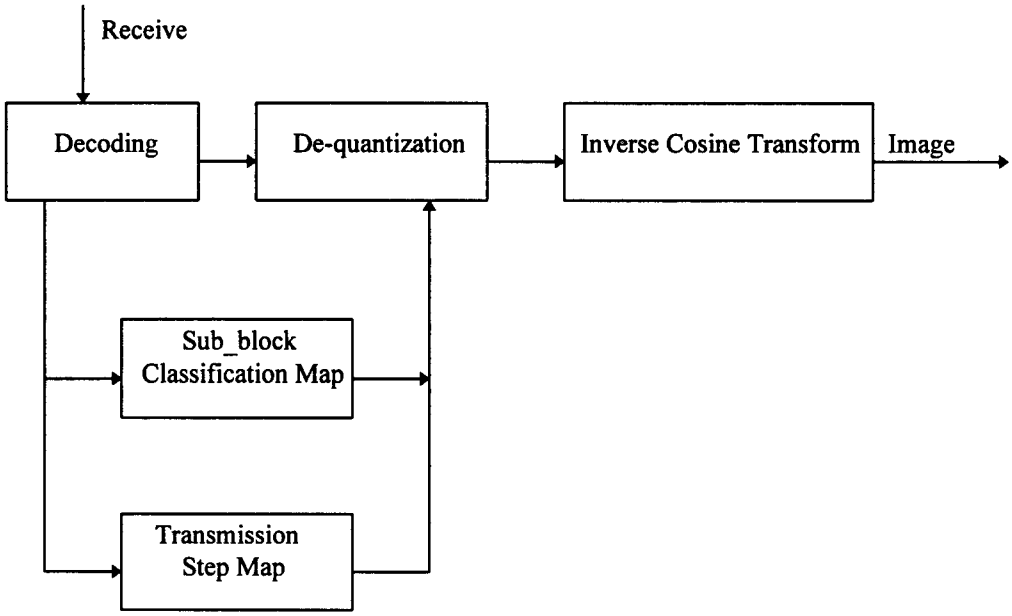


Fig. 7. Progressive Reconstruction System

5. Results

The software evolution of the proposed progressive transmission and reconstruction system is shown in Fig. 8. Simulations were carried out on Sun OS 5.4 under openwindows version 3.4. Modules have been developed in C language for Wavelet decomposition, Descrete Cosine Transform, Inverse Descrete Cosine Transform, Classification Maps, Transmission Step Maps, encoding and decoding. For experimental purposes an image of size 121x121 and a block size 11x11 is chosen. In each transmission step 121 samples are selected for transmission. Reconstructed images for the case of loss less coding are shown in fig. 9. As is evident from the figures, a quick recognition of the image can be made in between step0 and step9. An estimate of the mean square signal to noise ratio is computed by

$$SNR_{ms} = \frac{\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} \hat{f}(i,j)^2}{\sum_{i=0,j=0}^{N-1}\sum^{N-1} [\hat{f}(i,j) - f(i,j)]^2} \qquad (17)$$

where the $\hat{f}(i,j)$ and $f(i,j)$ represent the estimated image and the original image respectively. NxN is the size of the image. The computed values of SNR_{ms} for certain reconstruction steps are shown in table I.

Table I

Reconstruction step	SNR_{ms}
0	6.6087
10	11.2125
20	11.6768
30	11.7099
40	11.8812
50	12.0100
60	12.1096
70	12.1820
80	12.2349
90	12.2724
100	12.2935
110	12.3040
120	12.3111

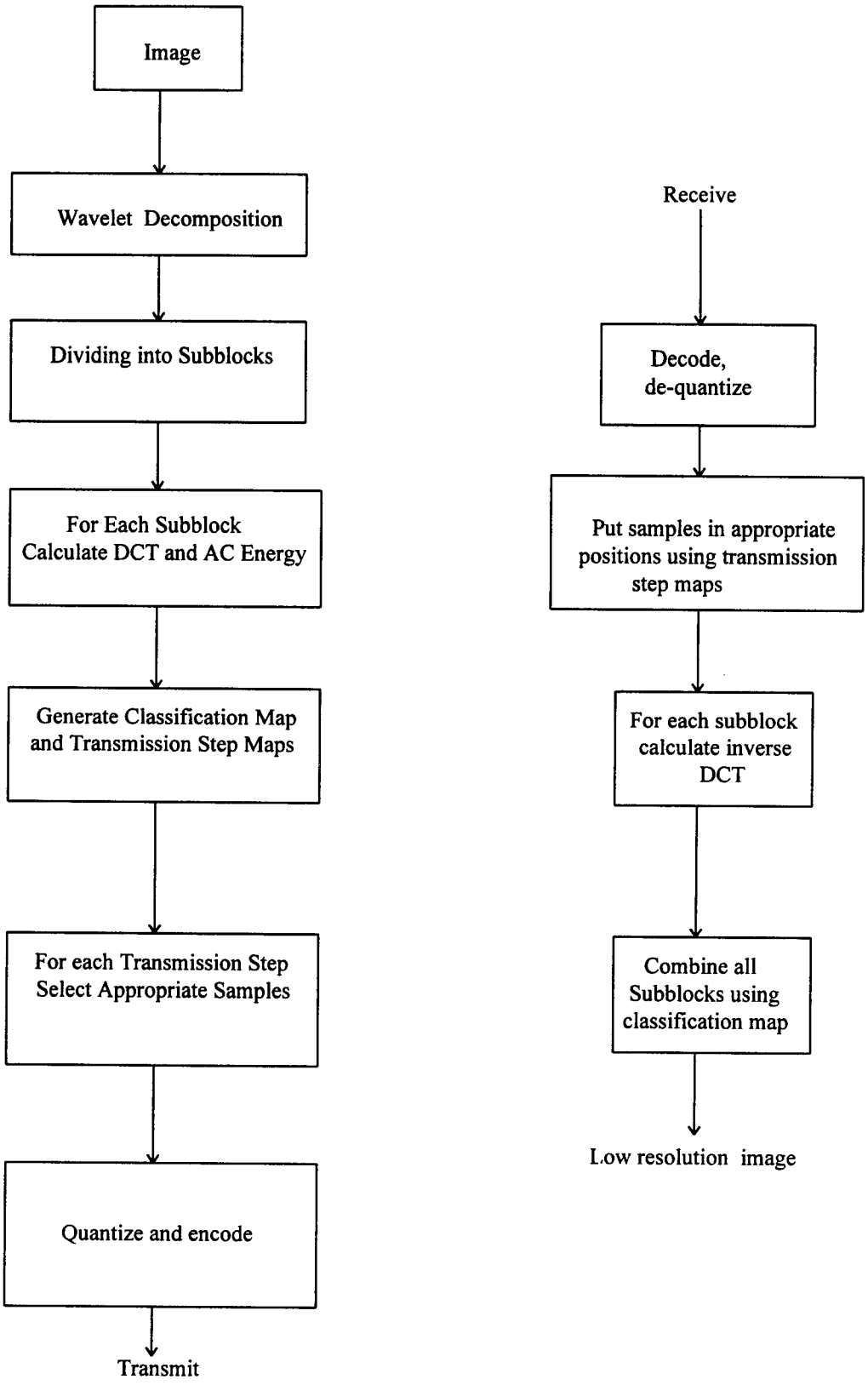


Fig. 8 Flow chart showing sequence of operations in progressive transmission and reconstruction

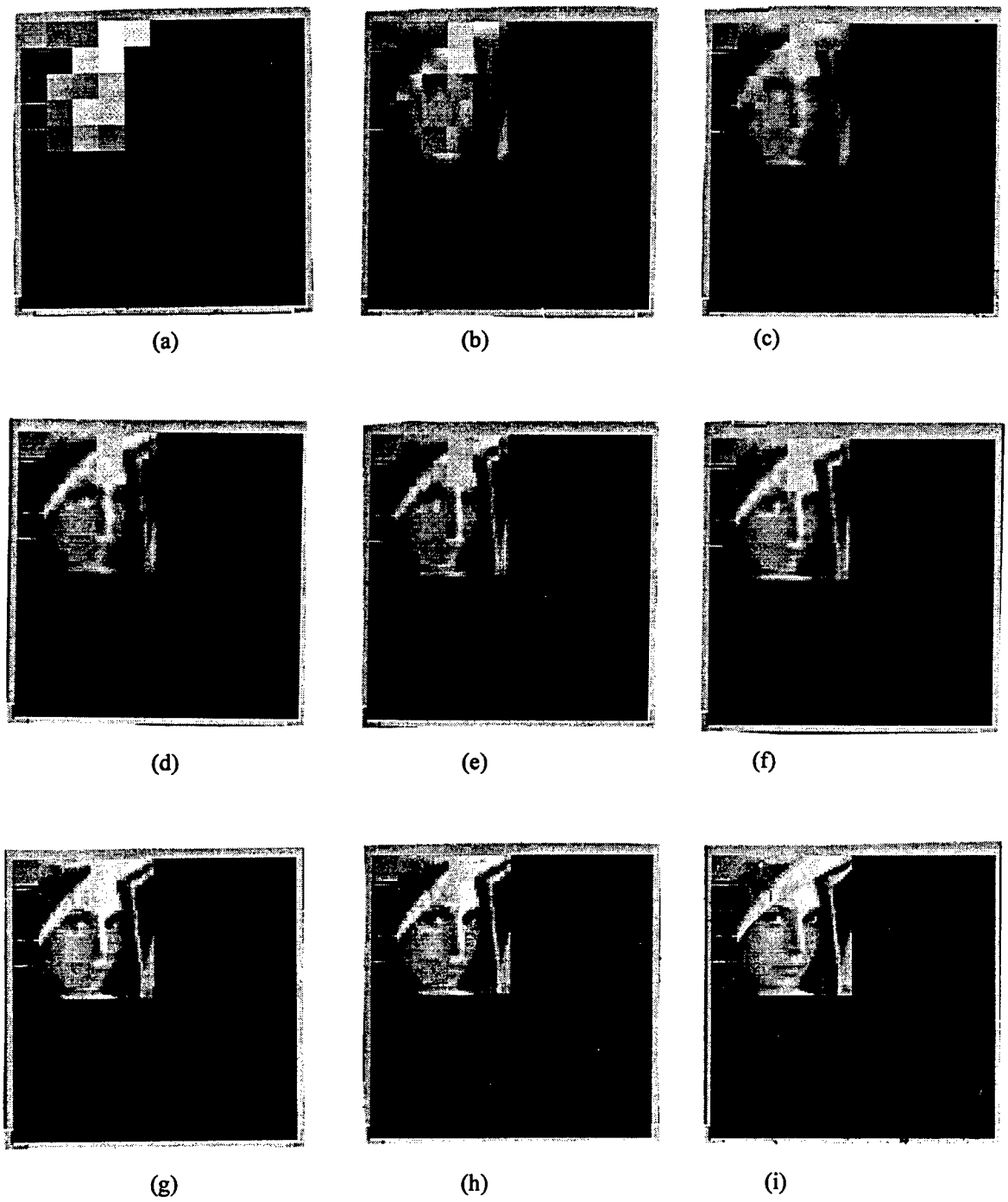


Fig. 9. Progressive reconstruction of images. Figures (a), (b), (c), (d), (e), (f), (g), (h) and (i) corresponds to step 0, step 1, step 2, step 3, step 4, step 5, step 8, step 9 and step 120 respectively.

6. Conclusions

We have demonstrated that the combined approach to progressive image transmission using wavelets and descrete cosine transform, enables a quick image recognition for the user who is browsing the image. Wavelets provide multiresolution analysis of an image and the importance of coefficients in each resolution for transmission is provided by the DCT method. Because wavelet decomposition produces lower resolution image and detail images, the user can see different resolutions of the image during browsing and can obtain nearly lossless reconstruction at each resolution by adding detail images. It is observed that most of the energy is concentrated in the lower resolution image, which is actually scattered in the original image. These coefficients get highest priority for transmission over detail coefficients enabling quick recognition of the image.

Our results for mean square signal to noise ratios in various reconstruction steps show that the recognition can be made in between step 0 and step 10. The improvement in mean square signal to noise ratio from step 11 to step 120 is marginal. It is because most of the high energy transform samples were received from step 0 to step 10 and the transform samples received after step 10 to step 120 were corresponding to mainly detail coefficients, which are usually very small in magnitude and adds to detail images.

References

1. Haar, "Zur theorie der orthogonalen Funktionensysteme," *Math. Ann.*, Vol. 69, pp. 331-371, 1910.
2. Gabor, "Theory of communications," *Journal of Insti. Elec. Eng.*, Vol. 93, pp. 429-457, 1946.
3. J. Morlet, G. Arens, E. Fourgeau, and D. Giard, "Wave Propagation and sampling theory I, II," *Geophysics*, Vol. 47, pp. 203-236, 1982.
4. I. Daubechies, "Orthonormal bases of compactly supported wavelets," *Comm. Pure and Appl. Math.*, Vol. 41, no. 7, pp. 909-996, 1988.
5. J. L. Starck and A. Bijaoui, "Filtering and deconvolution by the wavelet transform," *Signal Processing*, Vol. 35, pp. 195-211, 1994.
6. Olivier Rioul and Pierre Duhamel "Fast algorithms for discrete and continuous wavelet transforms," *IEEE Transactions on Information theory*, Vol. 38, pp. 569-583, No. 2, March 1992.
7. S. Mallat, " A theory of multiresolution signal decomposition: the wavelet representation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 11, no. 7, pp. 674-693, 1989.
8. N. Ahmed, T. Natarajan, and K. R. Rao, "Descrete cosine transform," *IEEE Trans. on Computers*, pp. 90-93, January 1974.

Investigation of Confined Optical Phonons for the Design of Si-Based Intersubband Lasers

Gang Sun
Assistant Professor
Engineering Program/Physics Department

University of Massachusetts at Boston
100 Morrissey Blvd.
Boston, MA 02125

Final Report for:
Summer Faculty Research Extension Program
Rome Laboratory
Hanscom Air Force Base

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC

and

Rome Laboratory

December 1996

Investigation of Confined Optical Phonons for the Design of Si-Based Intersubband Lasers

Gang Sun
Assistant Professor
Engineering Program/Physics Department
University of Massachusetts at Boston

ABSTRACT

The confinement of optical modes of vibrations in a superlattice consisting of polar and nonpolar materials is described by a continuum model. Specifically, the structure under investigation is the Si/ZnS superlattice. Optical phonon modes in Si and ZnS layers are totally confined within their respective layers since both layers can be treated as infinitely rigid with respect to the other layer. Since there are no associated electric fields with nonpolar optical phonons in Si layers, only mechanical boundary condition needs to be satisfied for these nonpolar optical modes at the Si-ZnS interface. The optical phonons in Si layers can be described by guided modes consisting of an uncoupled s-TO mode and a hybrid of LO and p-TO modes with no interface modes. In ZnS layers, a continuum model hybridizing the LO, TO and IP modes is necessary to satisfy both the mechanical and electrostatic boundary condition at the heterointerface. A numerical procedure is provided to determine the common frequency between LO, TO, and IP modes. Analytical expressions are obtained for the ionic displacement and associated electric field as well as scalar and vector potentials. These expressions can be employed directly in calculating the carrier interaction with optical phonons in the superlattice.

Investigation of Confined Optical Phonons for the Design of Si-Based Intersubband Lasers

Gang Sun

I. INTRODUCTION

With the demonstration of the InGaAs/AlInAs intersubband quantum cascade laser at $\lambda = 4.2\mu m$[1, 2], there has been interest in the possible utilization of silicon as the optically active material because of its integrability in advanced silicon microelectronics[3, 4]. In addition, there is interest in moving the lasing from the far and midinfrared range to the near infrared optical communication wavelengths, $\lambda = 1.3 \sim 1.55\mu m$[5]. Since the latter wavelength corresponds to a photon energy of $800meV$, the $Si_{1-x}Ge_x$/Si heterosystem is inadequate, since a maximum practical valence band offset of only of the order of 500 meV can be obtained for $x = 0.5 \sim 0.6$, for Si layers sufficiently thin not to exceed the critical thickness. Therefore, alternate large bandgap, nearly lattice matched, barrier materials must be sought with sufficiently large band offsets with respect to silicon. Possible candidates include ZnS, CaF_2, SiO_2 or the Si/SiO_2 superlattice, and γ-Al_2O_3, among others[5]-[7].

The Si/ZnS heterosystem has received the most attention as current advances in epitaxy technology have allowed the growth of heterostructures consisting of polar and nonpolar materials[8, 9]. The lattice mismatch of ZnS with respect to Si is only 0.3%. The valence band offset has been predicted theoretically and experimentally[10]-[13]. Values range between 700 and 1900 meV, sufficiently large to give intersubband energy differences in the desired range. Growth of ZnS upon Si and Si upon ZnS have been demonstrated[9], with the use of an As monolayer to satisfy the local bonding requirements although the affect of the monolayer on the offsets has not been determined.

The possibility of population inversion and the operation of the intersubband laser depend critically on the lifetimes of the involved subbands. The subband lifetimes in turn are determined by nonradiative phonon scattering processes. The purpose of the present paper is to study the optical phonon modes in the Si/ZnS system since their interaction with carriers is considered to be dominant in the phonon scattering processes. This combination of materials is new, since it consists of both a nonpolar and polar semiconductor. Previous studies in carrier scattering by confined optical phonons in heterostructures have been focused only on one type of phonons, either polar[14]-[22] and nonpolar[23]-[26]. In the current situation involving both polar and nonpolar materials, carrier scattering by both types of phonons needs to be considered. To the best of our knowledge, there has not been any reported investigation on this mixed nature of optical phonons, their confinement effect, and their interaction with carriers in a heterostructure. In this paper, we will present a theoretical study based on the macroscopic continuum model to describe the confined optical phonon modes in a heterostructure consisting of polar and nonpolar materials. The ultimate intersubband

laser design will likely consist of many periods, each of which will consist of more than one Si quantum wells coupled by ZnS barriers, with each period engineered to achieve population inversion. However, in this initial investigation, we will consider only a simple superlattice consisting of alternating layers of Si and ZnS. The results of this study will provide the basis for the more complex structure described above.

As described below in greater detail, since the optical dispersions (frequency versus wavevector) of the silicon (Si) and zinc sulphide (ZnS) have no overlap, the optical phonons are assumed to be totally confined in both materials. In the silicon layers, a continuum model with double hybridization of the longitudinal optical (LO) and transverse optical (TO) modes is used to describe the vibration patterns of the guided modes[23]. The only boundary condition that needs to be satisfied in the Si layers is the vanishing of the displacements at the Si-ZnS interface, since the ZnS layers can be considered as infinitely rigid with respect to the vibrations of the Si layer. Hence, there is no interface mode in the Si layers. The situation on the ZnS layers is more complex. Following the work by Ridley[15, 16], here a continuum model is employed with hybridization of the optical LO, TO, and interface polariton (IP) modes needed to satisfy both the mechanical and electrostatic boundary conditions at the interfaces. Specifically, the electrostatic boundary conditions are the continuity of E_x, the electric field parallel to the interface, and the continuity of D_z, the displacement field normal to the interface. The mechanical boundary condition is again the vanishing of the optical displacements since the Si layers can be considered as infinitely rigid with respect to the vibrations of the ZnS layers.

Our current work provides a complete set of analytical expressions for the optical phonon dispersion relations, optical displacements, and associated scalar and vector potentials. These expressions can be used directly in calculating the interaction of carriers with the confined optical phonons.

II. Mode Patterns and Dispersion Relationship

A continuum model for the optical modes in the Si/ZnS superlattice is employed. Both mechanical and electrical boundary conditions are satisfied at the heterointerfaces. Since the optical dispersion relations (frequency versus phonon wavevector) in the two bulk materials have no overlap, the phonons are taken to be confined in their respective materials. For the Si layers, the continuum model for optical phonons in nonpolar materials[23, 25] is used. Here double hybridization of the LO (longitudinal optical) and TO (transverse optical) modes is used to give the vibration patterns of the guided modes. Since the ZnS layers are infinitely rigid with respect to the vibrations of the Si layers, only the mechanical boundary condition, the vanishing of the displacements at the interfaces, has to be satisfied.

For the polar ZnS layers, an alternate continuum model developed by Ridley and coworkers[15, 16] is employed. The situation is more complex than for nonpolar materials. Here, in order to satisfy both the electrostatic and mechanical boundary

conditions, an intermixing of confined LO, TO, and IP (interface polariton) modes is needed. The boundary conditions which must be satisfied are (1) the continuity of E_x, the component of electric field parallel to the interface, (2) the continuity of D_z, the component of the displacement vector normal to the interface, and (3) the vanishing of the vector displacement u at the interface.

A. Modes in Si Layers

As discussed above, since the ZnS layers can be treated as infinitely rigid, the boundary condition to be satisfied in the Si layers is the vanishing of the ionic displacement of all confined vibration modes. This is an assumption of strict confinement yielding only the guided modes. As pointed out in the continuum theory[23], the ionic displacement of confined vibrations has two components: one is the hybrid of the LO and p-polarized TO (p-TO) modes, and other is the uncoupled s-polarized TO (s-TO) mode. These modes are defined as follows: If we consider a (x, z) plane containing the normal to the layers and the phonon wavevector $\mathbf{Q}$, then

$$\mathbf{Q} = q_x \hat{e_x} + q_z \hat{e_z} \tag{1}$$

where $\hat{e_x}$ and $\hat{e_z}$ are unit vectors. The p-TO mode has its displacements normal to $\mathbf{Q}$ and in the plane, while the s-TO displacements are normal to $\mathbf{Q}$ and perpendicular to the plane ($||\hat{e_y}$).

The form of the ionic displacement, scalar, and vector potentials in one superlattice period differs from that in a neighboring period only by a phase factor proportional to the Bloch superlattice wavevector q_{SL}. Their expressions given below are obtained by taking $q_{SL} = 0$. A description of the s-TO mode is

$$u_y = e^{iq_x x}(A_{s-TO}e^{iq_z z} + B_{s-TO}e^{-iq_z z}), \tag{2}$$

while the hybrid of the LO and p-TO modes is given by

$$\begin{aligned} u_x &= e^{iq_x x}[q_x(A_{LO}e^{iq_L z} + B_{LO}e^{-iq_L z}) + q_T(A_{p-TO}e^{iq_T z} + B_{p-TO}e^{-iq_T z})], \\ u_z &= e^{iq_x x}[q_L(A_{LO}e^{iq_L z} - B_{LO}e^{-iq_L z}) - q_x(A_{p-TO}e^{iq_T z} - B_{p-TO}e^{-iq_T z})]. \end{aligned} \tag{3}$$

which are confined within the Si layer with a width of d_{Si}, $0 < z < d_{Si}$. The z-components of the LO and TO wavevector have been distinguished by q_L and q_T, respectively.

Since the LO and TO modes must have the same frequency to be effectively coupled, we must satisfy the condition

$$\omega^2 = \omega_0^2 - \beta_L^2(q_x^2 + q_L^2) = \omega_0^2 - \beta_T^2(q_x^2 + q_T^2), \tag{4}$$

where β_L and β_T are the velocities of LO and TO dispersions in Si, respectively.

Using the boundary condition that $\mathbf{u} = 0$ at the interfaces gives for the s-TO mode

$$u_y = Ae^{iq_x x}\sin(q_z z), \qquad \text{with } q_z = \frac{n\pi}{d_{Si}} \tag{5}$$

where $n = 1, 2, \cdots$ and A is a mode coefficient. This mode does not mix with other modes.

The hybrid LO and p-TO modes admit two classes of solutions. The 'sine' solution is

$$\begin{aligned} u_x &= 2Be^{iq_x x} q_x [\cos(q_L z) - \cos(q_T z)], \\ u_z &= 2iBe^{ik_x x} [q_L \sin(q_L z) + \frac{q_x^2}{q_T} \sin(q_T z)], \end{aligned} \tag{6}$$

and the 'cosine' solution is

$$\begin{aligned} u_x &= 2iBe^{iq_x x} [q_x \sin(q_L z) + \frac{q_L q_T}{q_x} \sin(q_T z)], \\ u_z &= 2Be^{iq_x x} q_L [\cos(q_L z) - \cos(q_T z)] \end{aligned} \tag{7}$$

where

$$q_L = \frac{n_L \pi}{d_{Si}} \quad \text{and} \quad q_T = \frac{n_T \pi}{d_{Si}}, \tag{8}$$

where $n_L = 1, 2, \cdots$, $n_T = 3, 4, \cdots$, $n_T - n_L = 2, 4, 6, \cdots$, and B is a mode coefficient. No interface modes exist in the Si layer because of the boundary condition $\mathbf{u} = 0$.

B. Modes in ZnS Layers

The boundary conditions are the continuity of E_x, D_z, and the vanishing of $\mathbf{u}$ at the interfaces. These conditions can be satisfied by a unique linear combination of LO, TO, and IP modes with common frequency and common in-plane wavevector q_x,

$$\mathbf{u} = \mathbf{u}_{LO} + \mathbf{u}_{TO} + \mathbf{u}_{IP} \tag{9}$$

We will use this hybrid expression to calculate the electrical interaction with carriers which is considerably stronger than the optical deformation potential interaction. We need consider only the displacements u_x and u_z, since u_y associated with the s-TO mode has no related electric field and therefore does not interact with carriers electrically. Once again, the expressions are obtained by taking the Bloch superlattice wavevector $q_{SL} = 0$.

For the LO mode, the ionic displacement

$$\begin{aligned} u_x &= e^{i(q_x x - \omega t)} q_x (A_L e^{iq_L z} + B_L e^{-iq_L z}), \\ u_z &= e^{i(q_x x - \omega t)} q_L (A_L e^{iq_L z} - B_L e^{-iq_L z}) \end{aligned} \tag{10}$$

which is confined within the ZnS layer with a width of d_{ZnS}, $-d_{ZnS}/2 < z < d_{ZnS}/2$

The associated electric fields are

$$E_x = -\rho_o u_x, \qquad E_z = -\rho_o u_z, \tag{11}$$

where

$$\rho_o = \frac{e^*}{\epsilon_o \Omega}, \tag{12}$$

with the effective ionic charge

$$e^{*2} = M\Omega\omega_{LO}^2\epsilon_o^2(\frac{1}{\epsilon_\infty} - \frac{1}{\epsilon_s}) \tag{13}$$

where M is the reduced mass, ϵ_o is the permittivity of free space, ϵ_∞, ϵ_s are the high-frequency and static permittivities, and Ω is the volume of primitive unit cell. The scalar potential ϕ associated with the electric field $\mathbf{E} = -\nabla\phi$ is in turn given as

$$\phi = -i\rho_o e^{i(q_x x - \omega t)}(A_L e^{iq_L z} + B_L e^{-iq_L z}), \tag{14}$$

For the TO mode

$$\begin{aligned} u_x &= e^{i(q_x x-\omega t)} q_T (A_T e^{iq_T z} + B_T e^{-iq_T z}), \\ u_z &= -e^{i(q_x x-\omega t)} q_L (A_T e^{iq_T z} - B_T e^{-iq_T z}) \end{aligned} \tag{15}$$

The electric fields associated with this mode are negligible.

For the IP mode

$$\begin{aligned} u_x &= e^{i(q_x x-\omega t)} q_p (A_P e^{iq_p z} + B_P e^{-iq_p z}), \\ u_z &= ie^{i(q_x x-\omega t)} q_p (A_P e^{iq_p z} - B_P e^{-iq_p z}) \end{aligned} \tag{16}$$

The associated electric fields are

$$E_x = -\rho_p u_x, \qquad E_z = -\rho_p u_z, \tag{17}$$

where

$$\rho_p = \rho_o \frac{\omega^2 - \omega_{TO}^2}{\omega_{LO}^2 - \omega_{TO}^2}. \tag{18}$$

The electric fields associated with the interface modes propogate into the Si layers although they are treated as infinitely rigid and do not contain ZnS ionic displacement.

Being a transverse electromagnetic wave, there is a vector potential $\mathbf{A}$ associated with the electric field $\mathbf{E} = -\partial\mathbf{A}/\partial t$. Within the ZnS layers,

$$\begin{aligned} A_x &= i\frac{\rho_p}{\omega} e^{i(q_x x-\omega t)} q_p (A_P e^{iq_p z} + B_P e^{-iq_p z}), \\ A_z &= -\frac{\rho_p}{\omega} e^{i(q_x x-\omega t)} q_p (A_P e^{iq_p z} - B_P e^{-iq_p z}) \end{aligned} \tag{19}$$

While in the Si layers, a similar expression can be obtained with another set of mode coefficients, A_{p1} and B_{p1}.

Under the assumption of long wavelength waves and elastically isotropic medium, the requirement for common frequency gives the dispersion relationship,

$$\begin{aligned} \omega^2 &= \omega_{LO}^2 - v_L^2(q_x^2 + q_L^2) \\ &= \omega_{TO}^2 - v_T^2(q_x^2 + q_T^2) \\ &= \frac{c^2(q_x^2 + q_p^2)}{\epsilon(\omega)\mu_o} \end{aligned} \tag{20}$$

where v_L and v_T are velocities approximately equal to those of LA and TA modes in ZnS, respectively, c is the velocity of light in vacuum and μ_o is the permittivity of free space. In the above expressions, the frequency in the ZnS layers lies between the ZnS LO and TO zone center frequencies. Since $\omega_{TO} < \omega_{LO}$, in order for the TO frequency to be equal to a LO frequency q_T must be imaginary $q_T = iq_o$, corresponding to a TO interface mode. The modes which interact most strongly with carriers are those with frequencies near the LO branch. For these modes, the value of q_o is large, and we can take the approximation

$$\tanh(q_o d_{ZnS}) \approx 1. \tag{21}$$

In the unretarded limit ($c \to \infty$), $q_x^2 + q_p^2 \approx 0$ for the IP mode. Hence, $q_p \approx iq_x$.

Applying, at the two interfaces between layers Si and ZnS in a period of the superlattice, the conditions that u_x and u_z equal to zero along with the continuity of E_x and D_z, leads to eight simultaneous equations involving the eight unknown mode coefficients ($A_L, B_L; A_T, B_T; A_P, B_P$; and A_{P1}, B_{P1}). The following two ionic displacement mode patterns emerge for the hybrid in Eq.(9) taking the Bloch superlattice wavevector $q_{SL} = 0$ and the approximation $\tanh(q_o d_{ZnS}) \approx 1$. For the first type,

$$\begin{aligned}
u_x = \; & 2iBe^{iq_x x} q_x \{\sin(q_L z) \\
& -[1 - p_1 \tanh(q_x d_{ZnS}/2)] \sin(q_L d_{ZnS}/2) \frac{\sinh(q_o z)}{\sinh(q_o d_{ZnS}/2)} \\
& -p_1 \sin(q_L d_{ZnS}/2) \frac{\sinh(q_x z)}{\cosh(q_x d_{ZnS}/2)}\}, \\
u_z = \; & 2Be^{iq_x x} q_L \{\cos(q_L z) \\
& -\frac{q_x^2}{q_L q_o}[1 - p_1 \tanh(q_x d_{ZnS}/2)] \sin(q_L d_{ZnS}/2) \frac{\cosh(q_o z)}{\sinh(q_o d_{ZnS}/2)} \\
& -\frac{q_x}{q_L} p_1 \sin(q_L d_{ZnS}/2) \frac{\cosh(q_x z)}{\cosh(q_x d_{ZnS}/2)}\},
\end{aligned} \tag{22}$$

and for the second type,

$$\begin{aligned}
u_x = \; & 2Be^{iq_x x} q_x \{\cos(q_L z) \\
& -[1 - p_2 \coth(q_x d_{ZnS}/2)] \cos(q_L d_{ZnS}/2) \frac{\cosh(q_o z)}{\sinh(q_o d_{ZnS}/2)} \\
& -p_2 \cos(q_L d_{ZnS}/2) \frac{\cosh(q_x z)}{\sinh(q_x d_{ZnS}/2)}\}, \\
u_z = \; & 2iBe^{iq_x x} q_L \{\sin(q_L z) \\
& +\frac{q_x^2}{q_L q_o}[1 - p_2 \coth(q_x d_{ZnS}/2)] \cos(q_L d_{ZnS}/2) \frac{\sinh(q_o z)}{\sinh(q_o d_{ZnS}/2)} \\
& +\frac{q_x}{q_L} p_2 \cos(q_L d_{ZnS}/2) \frac{\sinh(q_x z)}{\sinh(q_x d_{ZnS}/2)}\},
\end{aligned} \tag{23}$$

where

$$
\begin{aligned}
p_1 &= -\frac{\alpha+\gamma}{2rsd}, \\
p_2 &= \frac{\gamma-\alpha}{2rsd}, \\
\alpha &= \sinh(q_x d_{Si})\cosh(q_x d_{ZnS}) + r\cosh(q_x d_{Si})\sinh(q_x d_{ZnS}), \\
\gamma &= \sinh(q_x d_{Si}) + r\sinh(q_x d_{ZnS}), \\
d &= 1 - Z\sinh(q_x d_{ZnS})\sinh(q_x d_{Si}) - \cosh(q_x d_{ZnS})\cosh(q_x d_{Si}), \\
Z &= \frac{1}{2}(r+\frac{1}{r}), \\
s &= \frac{\omega^2-\omega_{TO}^2}{\omega_{LO}^2-\omega_{TO}^2}, \\
r &= \frac{\epsilon_{p1}}{\epsilon_{p2}}
\end{aligned}
\tag{24}
$$

where ϵ_{p1} and ϵ_{p2} are the permittivities in Si and ZnS layers, respectively, with

$$
\epsilon_{p2} = \epsilon_\infty \frac{\omega^2-\omega_{LO}^2}{\omega^2-\omega_{TO}^2}. \tag{25}
$$

C. Dispersion Relationship

The phonon frequency in the ZnS layers is determined by the following set of equations;

$$
\begin{cases}
\omega^2 = \omega_o^2 - v_L^2(q_x^2+q_L^2), \\
\omega^2 = \omega_o^2 - v_T^2(q_x^2-q_o^2), \\
t_1 + t_2\cos(q_L d_{ZnS}) + t_3\sin(q_L d_{ZnS}) = 0
\end{cases}
\tag{26}
$$

where

$$
\begin{aligned}
t_1 = {} & 4p\sinh(q_x d_{Si}) + 4pr\sinh(q_x d_{ZnS}), \\
t_2 = {} & -4p\alpha, \\
t_3 = {} & 8p^2 r\sinh(q_x d_{ZnS})\sinh(q_x d_{Si}) - 4p^2\alpha^2 \\
& +4p^2 r^2\sinh^2(q_x d_{ZnS}) + 4p^2\sinh^2(q_x d_{Si}) + 1,
\end{aligned}
\tag{27}
$$

and

$$
p = \frac{q_x}{4q_L rsd}. \tag{28}
$$

The third equation in (26) is obtained from the requirement of a nonzero solution for the eight simultaneous equations discussed above, and Eq.(27) is arrived under the approximation, $\tanh(q_o d_{ZnS}) \approx 1$.

The numerical procedure for determining a phonon frequency is the following: given a value of q_x, we can determine those of t_1, t_2, and t_3 from Eq.(27). Then ω is scanned from ω_{TO} to ω_{LO}. For a given value of ω, q_L and q_o are obtained from the first two equations in (26). Those values are then substituted into the third equation in (26) to determine if the particular value of ω is a solution.

III. Scalar and Vector Potentials

The study of optical modes in the Si/ZnS superlattice will be ultimately applied to calculate the electrical interaction between the optical phonons and carriers in the superlattice. For this purpose, expressions for the scalar and vector potentials are essential in obtaining the electrical interaction Hamiltonian with respect to such an interaction,

$$H = -e\phi + \frac{e}{m}\mathbf{A}\cdot\mathbf{p}, \tag{29}$$

where $\mathbf{p}$ is the momentum operator, e and m are the free electron charge and mass, respectively. The scalar potential ϕ associated with the LO mode vanishes in Si layers where there is only the $\mathbf{A}\cdot\mathbf{p}$ interaction.

Associated with the two types of ionic displacement in Eqs.(22) and (23), the scalar potentials in ZnS layers are given as, for the first type,

$$\phi = \begin{cases} 2\rho_o B e^{iq_x x}\sin(q_L z_1) & |z_1| < \dfrac{d_{ZnS}}{2} \quad \text{ZnS layer,} \\ 0 & |z_2| < \dfrac{d_{Si}}{2} \quad \text{Si layer,} \end{cases} \tag{30}$$

and for the second type,

$$\phi = \begin{cases} -2i\rho_o B e^{iq_x x}\cos(q_L z_1) & |z_1| < \dfrac{d_{ZnS}}{2} \quad \text{ZnS layer,} \\ 0 & |z_2| < \dfrac{d_{Si}}{2} \quad \text{Si layer.} \end{cases} \tag{31}$$

Note that we have used two different coordinates z_1 and z_2 for layers ZnS and Si, respectively, with their origins placed at the centers of the respective layers.

The vector potentials can be obtained, for the first type,

$$A_x = \begin{cases} \dfrac{2s\rho_o q_x}{\omega} B e^{iq_x x} p_1 \sin(q_L d_{ZnS}/2)\dfrac{\sinh(q_x z_1)}{\cosh(q_x d_{ZnS}/2)} & |z_1| < \dfrac{d_{ZnS}}{2} \quad \text{ZnS layer,} \\ -\dfrac{4q_x\rho_o}{\omega d} B e^{iq_x x} V_1 \sinh(q_x z_2) & |z_2| < \dfrac{d_{Si}}{2} \quad \text{Si layer,} \end{cases} \tag{32}$$

$$A_z = \begin{cases} \dfrac{2is\rho_o q_x}{\omega} B e^{iq_x x} p_1 \sin(q_L d_{ZnS}/2)\dfrac{\cosh(q_x z_1)}{\cosh(q_x d_{ZnS}/2)} & |z_1| < \dfrac{d_{ZnS}}{2} \quad \text{ZnS layer,} \\ -\dfrac{4iq_x\rho_o}{\omega d} B e^{iq_x x} V_1 \cosh(q_x z_2) & |z_2| < \dfrac{d_{Si}}{2} \quad \text{Si layer,} \end{cases} \tag{33}$$

and for the second type,

$$A_x = \begin{cases} -\dfrac{2is\rho_o q_x}{\omega} B e^{iq_x x} p_2 \cos(q_L d_{ZnS}/2)\dfrac{\cosh(q_x z_1)}{\sinh(q_x d_{ZnS}/2)} & |z_1| < \dfrac{d_{ZnS}}{2} \quad \text{ZnS layer,} \\ -\dfrac{4iq_x\rho_o}{\omega d} B e^{iq_x x} V_2 \cosh(q_x z_2) & |z_2| < \dfrac{d_{Si}}{2} \quad \text{Si layer,} \end{cases} \tag{34}$$

$$A_z = \begin{cases} -\dfrac{2s\rho_o q_x}{\omega} B e^{iq_x x} p_2 \cos(q_L d_{ZnS}/2) \dfrac{\sinh(q_x z_1)}{\sinh(q_x d_{ZnS}/2)} & |z_1| < \dfrac{d_{ZnS}}{2} \quad \text{ZnS layer,} \\ -\dfrac{4q_x\rho_o}{\omega d} B e^{iq_x x} V_2 \sinh(q_x z_2) & |z_2| < \dfrac{d_{Si}}{2} \quad \text{Si layer,} \end{cases} \tag{35}$$

where

$$\begin{aligned} V_1 = \; & \sin(q_L d_{ZnS}/2)\cosh(q_x d_{ZnS}/2) \\ & [\cosh(q_x d_{ZnS}/2)\sinh(q_x d_{Si}/2) + r\sinh(q_x d_{ZnS}/2)\cosh(q_x d_{Si}/2)], \\ V_2 = \; & \cos(q_L d_{ZnS}/2)\sinh(q_x d_{ZnS}/2) \\ & [\sinh(q_x d_{ZnS}/2)\cosh(q_x d_{Si}/2) + r\cosh(q_x d_{ZnS}/2)\sinh(q_x d_{Si}/2)]. \end{aligned} \tag{36}$$

IV. Results and Discussion

A. Mode Patterns in Si Layers

The lowest s-TO mode pattern in Eq.(5) for $q_z = \pi/d_{Si}$ is shown in Fig.1(a) within a Si layer of $d_{Si} = 40\mathring{A}$, while the hybrid patterns of the lowest p-TO and LO modes with $q_L = \pi/d_{Si}$ and $q_T = 3\pi/d_{Si}$ are shown in Figs.1(b) and 3(c) for the 'sine' and 'cosine' solutions given in Eqs.(6) and (7), respectively within the same Si layer. The strict confinement which requires the vanishing of ionic displacements at the boundaries of Si layers is clearly demonstrated for both vibration modes.

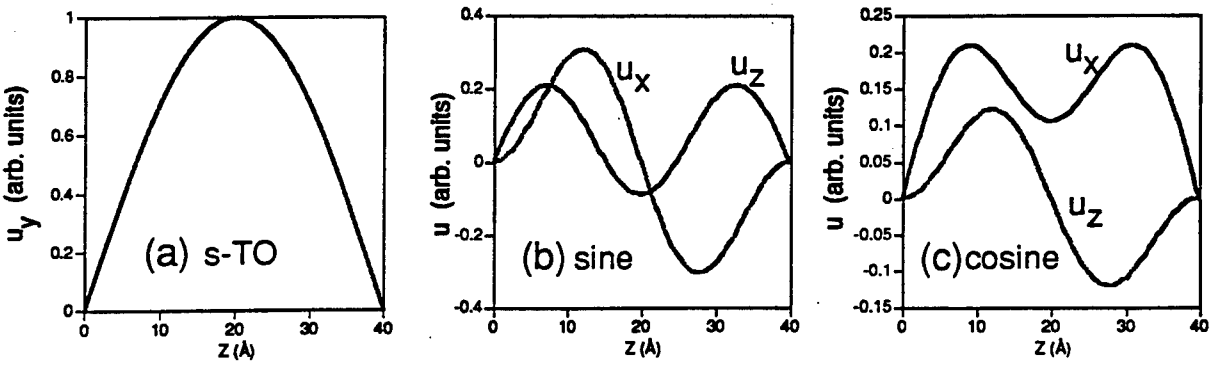


Figure 1. Vibration patterns in a Si layer with a width of $40\mathring{A}$ for (a) the guided s-TO mode, (b) the 'sine' solution, and (c) the 'cosine' solution of the guided p-TO and LO modes

B. Mode Patterns in ZnS layers

To illustrate the patterns of ionic displacements in the ZnS layers given in Eqs.(22) and (23), we need to first determine values for q_x, q_L, and q_o. To do so, we will follow the numerical procedure described in Section II(C) by arbitrarily fixing a value for the in-plane phonon wavevector $q_x = \pi/(10a_{ZnS})$, where a_{ZnS} being the lattice constant of ZnS. This choice of q_x satisfies the requirement for in-plane wavevectors to be considered

large enough to neglect the effect of retardation so that $q_p = iq_x$, and also at the same time small enough so that the quadratic dispersion assumed for the LO and TO modes in Eq.(20) is valid. In the event of calculating the carrier-optical phonon interaction, the value of q_x is actually determined by the conservation of in-plane momentum between the initial and final states of the scattering process. For a given value of q_x, typically, a set of hybridized modes can be obtained. Here, we show only the mode pattern with frequency close to ω_{LO}.

We obtained $\hbar\omega = 35.5meV$, $q_L = 0.31 \times 10^8/cm$ and $q_o = 0.98 \times 10^8/cm$. Fitting these values into Eqs.(22) and (23), we obtained Figs.2(a) and 2(b) showing the mode patterns of ionic displacement of both the first and second types, respectively in a ZnS layer of $d_{ZnS} = 20\mathring{A}$. It can be seen from Figs.2(a) and 2(b) that the mechanical boudary condition, vanishing of the ionic displacements at the interfaces of Si and ZnS layers, is satisfied.

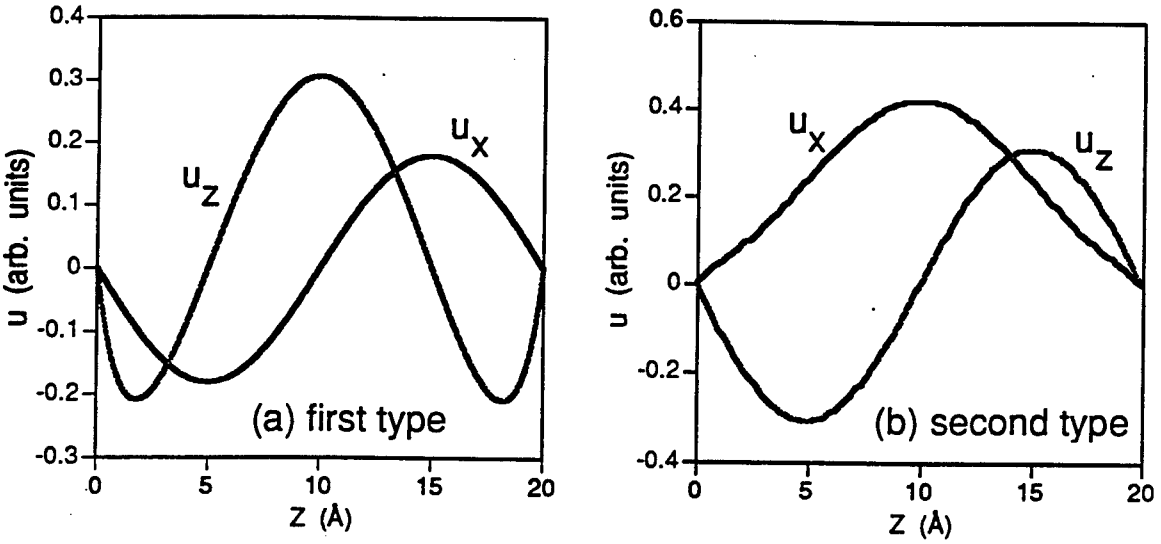


Figure 2. Vibration patterns in a ZnS layer with a width of $20\mathring{A}$ for (a) the first type and (b) the second type solutions of the hybridized LO, TO and IP modes.

C. Potential and Field Distributions in the Superlattice

The scalar potentials associated with the LO modes are strictly confined within the ZnS layers. Their distributions are shown in Fig.3 for the first and second types given in Eqs.(30) and (31) with $q_L = 0.31 \times 10^8/cm$, $d_{ZnS} = 20\mathring{A}$, respectively.

The vector potential associated with the IP modes are distributed in both Si and ZnS layers, even thought Si layers are treated as infinitely rigid and do not contain ZnS ionic displacements. The profiles for the two components of the vector potentials given in Eqs.(32-35) for the first and second types with $d_{Si} = 40\mathring{A}$, $d_{ZnS} = 20\mathring{A}$ are shown in Figs.4(a) and 4(b), respectively.

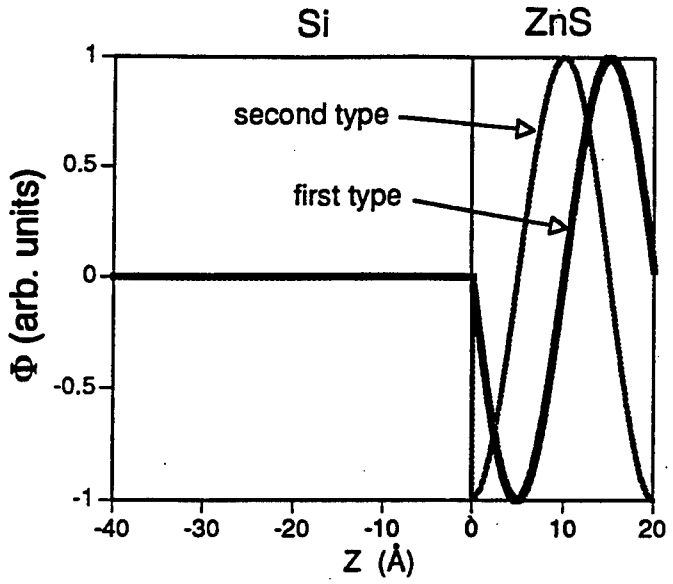


Figure 3. Scalar potential distribution associated with the LO modes in a period of the Si/ZnS superlattice with $d_{Si} = 40\mathring{A}$ and $d_{ZnS} = 20\mathring{A}$ for both the first and second types of the vibration modes.

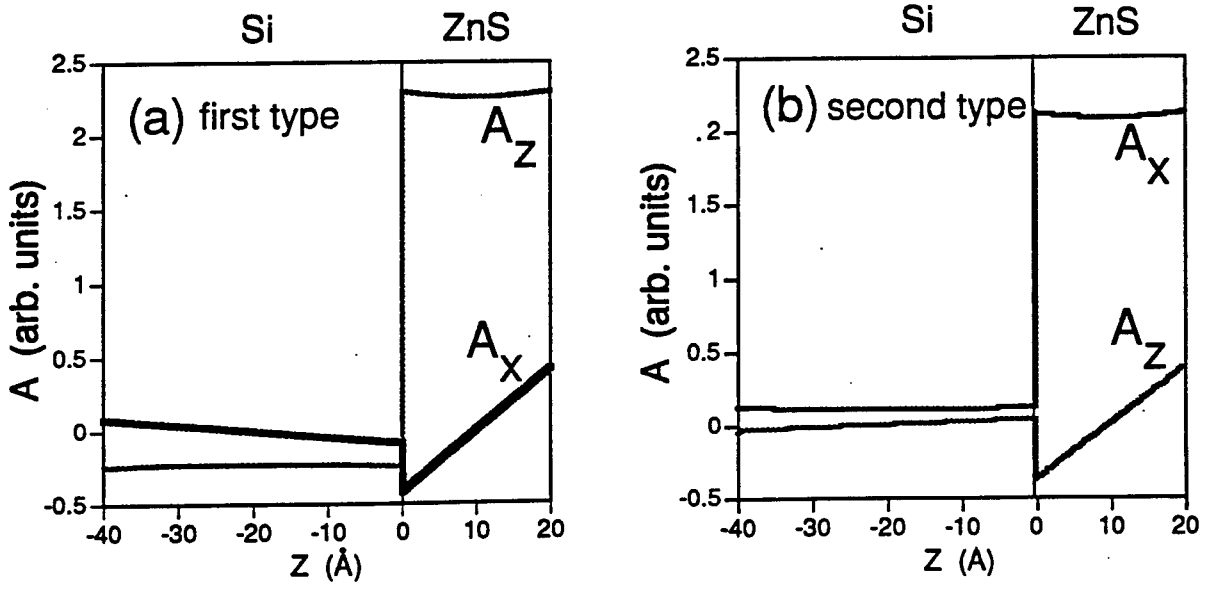


Figure 4. Vector potentials associated with the IP modes distributed in a period of the Si/ZnS superlattice with $d_{Si} = 40\mathring{A}$ and $d_{ZnS} = 20\mathring{A}$ for (a) the first type and (b) the second type of the vibration modes

It can be seen from Figs.3 and 4 that both scalar and vector potentials are not continuous across the interfaces. However, as pointed by Ridley[15], the energy of interaction with an electron traveling coherently with the optical phonon is continuous. The electric field can be obtained as

$$\mathbf{E} = -\nabla\phi - \frac{\partial \mathbf{A}}{\partial t}. \tag{37}$$

The continuity of E_x and $D_z = \epsilon(\omega)E_z$ implies that at boundaries,

$$\begin{aligned} \omega A_x|_{z_2=\pm d_{Si}/2} &= -q_x\phi|_{z_1=\mp d_{ZnS}/2} + \omega A_x|_{z_1=\mp d_{ZnS}/2}, \\ A_z|_{z_2=\pm d_{Si}/2} &= r A_z|_{z_1=\mp d_{ZnS}/2}, \end{aligned} \tag{38}$$

where A_{1x} and A_{1z} are x- and z-components of the vector potential in Si layers. The interaction in the Si layer is $e(A_{1x}v_x + A_{1z}v_z)$ and in the ZnS layer $e(-\phi + A_x v_x + A_z v_z)$, which are equal when the electron velocity $v_x = \omega/q_x$ and $v_z = 0$. Thus, the coherent interaction energy is continuous across the interfaces.

The electric field distributions for E_x and $\epsilon(\omega)$ in Si ($d_{Si} = 40\mathring{A}$) and ZnS ($d_{ZnS} = 40\mathring{A}$) layers are shown in Figs.5(a) and 5(b) for the first and second types, respectively. The continuity of E_x and D_z across the Si and ZnS interface according to Eq.(38) is clearly demonstrated.

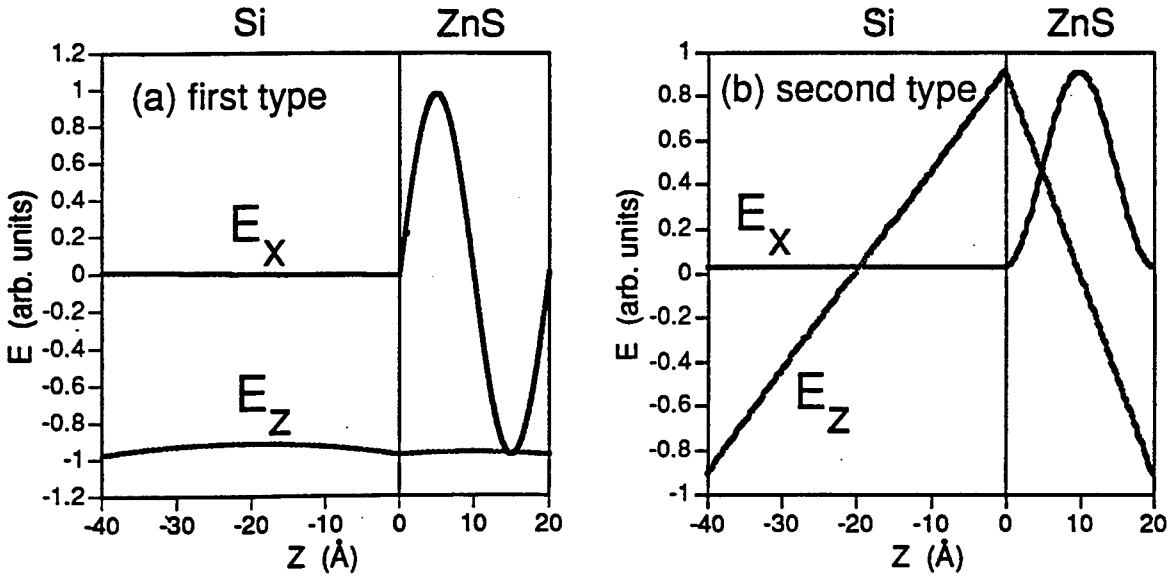


Figure 5. The field distributions, E_x and D_z, derived from the scalar and vector potentials, in a period of the Si/ZnS superlattice with $d_{Si} = 40\mathring{A}$ and $d_{ZnS} = 20\mathring{A}$ for (a) the first type and (b) the second type of the vibration modes.

V. Conclusions

We have provided an analytical model of optical modes in Si/ZnS superlattices consisting of polar and nonpolar optical phonons. In the Si layers, a continuum model with double hybridization of the LO and TO modes is used to described the vibration patterns. Since there is no electric field resulted from the nonpolar ionic displacements in Si layers, the only boundary condition that needs to be satisfied in the Si layers is the vanishing of the displacements at the Si-ZnS interface as the ZnS layers can be considered as infinitely rigid with respect to the vibrations of the Si layer. Due to this strict confinement, only guided modes emerge in the Si layers which consist of s-TO and coupled p-TO and LO modes, with no interface modes. These guided modes have been illustrated. Their interaction with carriers in the superlattice can be calculated through the optical deformation potential for Si. The interaction Hamiltonian can actually be obtained by taking the product of this potential with the normalized ionic displacement.

However, for the optical phonons in ZnS layers, we need to include the electrical interaction in calculating the carrier scattering by optical phonons, since there are electric fields associated with the polar optical vibrations. As a result, both mechanical and electrostatic boundary conditions needs to be satisfied in the interfaces. A continuum model employing a linear combination of LO, TO and IP (interface polariton) modes with a common frequency is used to describe the ionic displacements in ZnS layers. A numerical procedure for determining a phonon frequency is provided. This hybridized model is necessary to meet the simultaneous requirement on the mechanical and electrostatic boundary conditions. The mechanical boundary condition is again the vanishing of the optical displacements since Si layers can be considered as infinitely rigid with respect to the vibrations of the ZnS layers. The electrostatic boundary conditions are the continuity of the electric field parallel to the interface, and the continuity of the displacement field normal to the interface. Based on this set of boundary conditions, expressions are obtained for the ionic displacements in ZnS layers consisting of LO, TO, and IP modes. There are scalar and vector potentials associated with the LO and IP modes, respectively, but no electric field associated with the TO mode. The scalar potential and its associated electric field due to the LO mode are distributed only within the ZnS layers and are zero in the Si layers. But the vector potential and its associated electric field due to the IP mode have distributions in both ZnS and Si layers even though there is no ZnS ionic displacement mode in the Si layers. Examples of these mode characteristics have been demonstrated. Neither the scalar nor vector potential is continuous across the Si-ZnS interface, but the energy of coherent interaction with carriers is continuous due to the continuity of the electric field parallel to the interface. The analytical model for the confined optical modes consisting of polar and nonpolar optical phonons will be employed in calculating the carrier-phonon interaction to estimate the subband lifetimes in the Si/ZnS superlattices.

ACKNOWLEDGEMENTS

The author wishes to thank Dr. Richard A. Soref for his support during the period of this summer research program, without which this research accomplishment would not have been possible. The author would also like to acknowledge Dr. Lionel Friedman for many helpful discussions.

References

[1] J. Faist, F. Capasso, D. L. Sivco, A. L. Hutchinson, C. Sirtory, and A. Y. Cho, Science **264**, 553 (1994)

[2] J. Faist, F. Capasso, D. L. Sivco, A. L. Hutchinson, C. Sirtory, S. N. G. Chu, and A. Y. Cho, Appl. Phys. Lett. **65**, 2091 (1994)

[3] G. Sun, L. Friedman, and R.A. Soref, Appl. Phys. Lett. **66**, 3425 (1995)

[4] R. A. Soref, Proc. IEEE **81**, 1687 (1993)

[5] L. Friedman and R. A. Soref, IEEE Photonics Technology Letters **5**, 1200 (1993)

[6] L. J. Schowalter and R. W. Fathauer, CRC Critical Review **15**, 367 (1989)

[7] R. Tsu, Nature **364**, 19 (1993)

[8] M Yokoyama, K. I. Kashiro, and S. I. Ohta, J. Crystal Growth **81**, 73 (1987)

[9] X. Zhou and W. P. Kirk, Mat. Res. Soc. Symp. Proc. **318**, 207 (1994)

[10] E. G. Wang and C. S. Ting, Phys. Rev. B **51**, 9791 (1995)

[11] C. Maierhofer, S. Kulkarni, M. Alonso, T. Reich, and K. Horn, J. Vac. Sci. Technol. B **9**, 2238 (1991)

[12] M. Cardona and N. E. Christensen, J. Vac. Sci. Technol. B **6**, 1285 (1988)

[13] W. A. Harrison, J. Vac. Sci. Technol. **14**, 1016 (1977)

[14] B. K. Ridley, Phys. Rev. B **39**, 5282 (1989)

[15] B. K. Ridley, Phys. Rev. B **47**, 4592 (1993)

[16] M. P. Chamberlain, M Cardona, and B. K. Ridley, Phys. Rev. B **48**, 14356 (1993)

[17] N. C. Constantinou and B. K. Ridley, Phys. Rev. B **49**, 17065 (1994)

[18] B. K. Ridley, Appl. Phys. Lett. **66**, 3633 (1995)

[19] E. Molinari and A. Fasolino, Appl. Phys. Lett. **54**, 1220 (1989)

[20] L. register, Phys. Rev. B **45**, 8756 (1992)

[21] K. J. Nash, Phys. Rev. B **46**, 7723 (1992)

[22] N. Mori and T. Ando, Phys. Rev. B **40**, 6175 (1989)

[23] G. Sun and L. Friedman, Phys. Rev. B **53**, 3966 (1995)

[24] A. Fasolino, E. Molinari, and J.C. Mann, Phys. Rev. B **39**, 3923 (1989)

[25] B. K. Ridley, Phys. Rev. B **44**, 9002 (1991)

[26] S. C. Jain and W. Hayes, Semicond. Sci. Technol. **6**, 547 (1991)

Numerical Study of Bistatic Scattering From Land Surfaces at Grazing Incidence

James C. West
Associate Professor
School of Electrical Engineering

Oklahoma State University
Stillwater, OK

Final Report for:
Summer Research Extension Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base
Washington, D.C.

and

Rome Laboratory

December 1996

NUMERICAL STUDY OF BISTATIC SCATTERING FROM LAND SURFACES AT GRAZING INCIDENCE

James C. West
Associate Professor
School of Electrical and Computer Engineering
Oklahoma State University

Abstract

A numerical study has been performed to examine the effects of surface self-shadowing on the electromagnetic scatter from rough dielectric interfaces. A hybrid numerical technique combining the moment method and geometrical theory of diffraction was used in the numerical calculations. This technique was first extended to be applicable to general dielectric media as well as perfectly conducting and high loss, high permittivity surfaces. The numerical calculations show that the shadowing becomes stronger at vertical polarization and weaker at horizontal polarization as the magnitude of the dielectric constant of the scatterer becomes smaller. However, deeply shadowed roughness still has a greater contribution to the total backscatter at vertical polarization, even with a dielectric constant as small as 3. Weakly shadowed roughness can contribute to the backscatter at either vertical or horizontal polarization. The accuracy of the two-scale scattering model is not significantly improved by including correction for shadowing.

I. INTRODUCTION

Electromagnetic scattering from rough surfaces is affected by surface self-shadowing when the illumination grazing angle is small (large incidence angles). Traditional rough-surface scattering theories such as the Kirchhoff approximation (KA) [1], small-perturbation method (SPM) [2], and the two-scale model [3] that have proven accurate at moderate grazing angles do not directly account for shadowing, and therefore fail at the smallest grazing angles. Several attempts have been made to account for the shadowing by introducing a "shadowing function" that reduces the total scattering predicted by the traditional methods [4, 5, 6, 7]. These shadow functions are usually derived from a high frequency, geometrical optics representation of the shadowing, and assume that surface features within the shadowed regions do not significantly contribute to the scattered field.

The validity of the shadowing function approach under certain conditions has recently been examined. Using analytical approaches, Barrick [8] and Holliday *et al.*[9] both showed that the surface currents within a shadowed region can be quite high when the illumination is vertically polarized, even when the shadowing obstacle is electromagnetically large, and therefore may contribute to the overall scatter. The shadow-region currents were much smaller at horizontal polarization. West [10] used a numerical approach to show that deeply shadowed small-scale roughness can yield significant backscatter at vertical polarization, but was not important at horizontal polarization. These studies were limited to perfectly conducting or high loss, high dielectric constant surfaces. On the other hand, Mockapetris [11] experimentally showed that a good prediction of horizontally polarized bistatic scattering from terrain was obtained when surface self-shadowing was ignored, in apparent contradiction to studies described above. A possible reason for this discrepancy is that the experiment was performed at the White Sands Missile Range, where the dry soil is more properly modeled as a lossy dielectric rather than a perfect conductor.

Numerical calculation has often been used to examine the validity of scattering models under various conditions. One of the most popular numerical approaches used has been the moment method [12, 13, 14, 15]. The primary advantage of the moment method is that the exact scattering equations are numerically solved to yield the induced currents on the surface. The solution will therefore in theory converge to the exact solution (in the absence of numerical errors) as the fineness of the surface representation is increased, thereby giving an excellent benchmark with which to compare the scattering model predictions. (In fact, moment method calculations are often referred as "exact" [14].) The primary disadvantage of the approach is that

it yields a dense system of linear equations that must be solved, and the order of the system increases with either increasing fineness of the numerical description or with increasing electromagnetic dimensions of the scatterer itself. This limits the size of the scatterer that can be treated with the moment method, due to both finite computer memory and processing resources, and the increasing round-off errors with increasing system size. When applied to rough surface scattering, the modeled surface must therefore be truncated, giving non-physical edges that can lead to diffraction that affects the calculated scattering.

At moderate grazing angles these "edge effects" have traditionally been reduced to insignificant levels by applying an illumination weighting function that smoothly reduces the incident field to zero at the edges, thus eliminating the direct-illumination diffraction. Unfortunately, this approach cannot be efficiently applied at small grazing angles. Thorsos [14] showed that an electromagnetically valid weighting function becomes a narrow pencil beam at small grazing, leading to unrealistic illumination of the surface feature unless quite long surfaces are modeled. West [10] overcame this limitation by using a hybrid numerical technique that combines the moment method (MM) with the geometrical theory of diffraction (GTD) to allow the modeled surface to be extended to infinity, eliminating the artificial edges. Single moment method basis functions derived from GTD were used on the infinite extensions, thereby minimizing the computational expense of using an infinite extension [16]. The technique was recently extended by West and Sturm [17] to apply to high loss, high dielectric constant scattering media using impedance boundary conditions.

We have performed a numerical study to investigate the validity of geometrical-optics based shadowing theory when applied to surfaces whose dielectric properties approximate that of soil of differing moisture content. A numerical approach based on the hybrid MM/GTD technique was used to find the scattering from two-scale rough surfaces that allowed the shadowing to be controlled. The numerical technique was first extended to allow the application to low-loss, homogeneous dielectric scattering media. It was then applied to two-scale rough interfaces that had random small-scale roughness superimposed on a deterministic large-scale displacement. The scattering with small-scale roughness both included and not included in the shadowed regions was calculated and compared, giving an estimate of the importance of the shadow-region roughness. Although the original intent of the project was to examine bistatic as well as monostatic scattering, the extension of the numerical technique proved quite time consuming, and only the monostatic calculations have been completed.

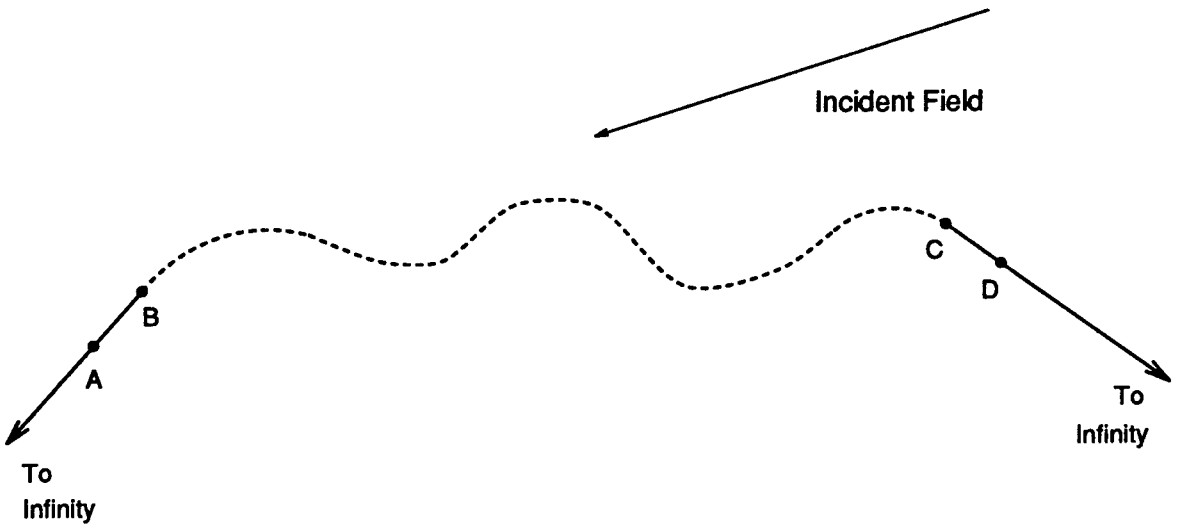


Figure 1: Arbitrary scattering surface.

II. NUMERICAL TECHNIQUE

Three different forms of the hybrid MM/GTD numerical scattering technique are used in this work, developed for perfectly conducting, high dielectric constant and high loss, and low dielectric/low loss media. Detailed overviews of each are given here.

A. *Perfectly Conducting Surfaces*

Application of the hybrid MM/GTD numerical technique to scattering from perfectly conducting surfaces is described in [10]. Adapted from the technique described by Burnside *et al.*[16], it is similar to the standard moment method in that scattering from the surface is found by first numerically solving an integro-differential equation to yield the surface current. In both methods, the unknown surface current is represented as a summation of known basis functions. The weighting coefficients associated with each basis function that give the "best" approximate solution are obtained using the moment method. The primary difference between the two techniques is that the hybrid approach uses *a priori* knowledge of the current, obtained from GTD, to define well behaved basis functions and additional source terms that allow the treatment of special infinitely long surfaces. The artificial edge effects introduced in the standard MM are thereby avoided. The surface current is then radiated to yield the scattered field.

For one-dimensionally rough surfaces of the type shown in Figure 1 considered here, vertically polarized

scattering is best described by the magnetic field integral equation (MFIE) [18]:

$$\begin{aligned} -H_z^i(l^-) &= -j\frac{k}{4}\int_S J_l(l')\,(\hat{\mathbf{n}}'\cdot\rho')\,\mathrm{H}_1^{(2)}(\beta|\rho^- - \rho'|)\,dl' \\ &= H_z^s[\rho^-, J_l(l)] \\ &= L_1^-[J_l(l)], \end{aligned} \tag{1}$$

where $H_z^i(l^-)$ is the incident, z-directed magnetic field at the surface, l^- is the arc length along the scattering surface, $J_l(l)$ is the unknown surface current to be found (oriented along the arc of the interface, referenced to the direction of integration), k is the free space wave number, ρ^- is the position vector of the observation point (corresponding to l^-), ρ' is the position vector of the source point (corresponding to l'), $\hat{\mathbf{n}}'$ is the normal unit vector at the source point, and $\mathrm{H}_1^{(2)}$ is the first-order Hankel function of the second type. The $^-$ superscript indicates that equation (1) is evaluated just inside the surface of the scatterer. $H_z^s[J_l(l), l']$ indicates that the right hand side of equation (1) gives the field scattered by the surface current. Setting it equal to the negative of the incident field forces the total field inside the perfectly conducting surface to zero. (The perfectly conducting boundary conditions can be used to recast (1) to the more familiar form [18]

$$H_z^i(l) = 0.5J_l(l) + j\frac{\beta}{4}\int_S J_l(l')\,(\hat{\mathbf{n}}'\cdot\rho')\,\mathrm{H}_1^{(2)}(\beta|\rho^- - \rho'|)\,dl' \tag{2}$$

where l indicates the evaluation is exactly on the surface and the integration is the principal value integral that avoids the singularity at $l' = l$.)

Horizontally polarized scattering is more easily treated by the electric field integral equation (EFIE):

$$\begin{aligned} -E_z^i(l^-) &= \frac{k\eta_0}{4}\int_S J_z(l')\mathrm{H}_0^{(2)}(k|\rho^- - \rho'|)\,dl' \\ &= E_z^s[\rho^-, J_z(l)] \\ &= L_2^-[J_z(l)], \end{aligned} \tag{3}$$

where $E_z^i(l^-)$ is the incident, z-directed electric field immediately below the surface, η_0 is the intrinsic wave impedance of free space. Again, the right hand side gives the field scattered by the current, and setting it equal to the negative incident field gives zero field within the scatterer.

The MFIE and EFIE can be written in the single notation

$$-F_z^i(l^-) = L_N^-[J_u(l)], \tag{4}$$

where F is E or H, N is 1 or 2, and u is z or l. In the standard moment method, the infinite integrations in equations (1) and (3) are truncated to be over a finite surface arc length. The current on the modeled length is then divided into a weighted summation of adjacent pulse basis functions, and the moment method is used to find the associated weighting coefficients. It is the truncation of the integrations that lead to the non-physical edge effects.

The hybrid MM/GTD technique is applied to one-dimensionally rough surfaces of the form shown in Figure 1. The dashed section of the surface represents the actual rough surface while the solid line represents infinitely long, planar extensions. The extensions are chosen such that all points on the actual surface are shadowed from all points on the extension (except of course at the intersection points B and C). Because the surface is arbitrary, little is known initially about the current between points A and D. Thus, the current in this region is described using standard MM pulse basis functions with impulse testing functions (yielding point matching) centered on the basis functions.

Since the extensions are shadowed from the arbitrary surface points, the fields at the surface of the extensions can be entirely described as the sum of a field diffracted from point B or C plus the geometrical optical (GO) incident and reflected fields:

$$F^t = F^i + F^s = F^{GO} + F^d, \tag{5}$$

where F^t is the total field, F^i is the incident field, F^s is the scattered field, F^{GO} is the geometrical optics incident and reflected fields, and F^d is the diffracted field. The current on the extension is obtained by applying the surface boundary conditions to equation (5), yielding the physical optics current associated with the GO fields plus an additional current component associated with the diffracted field (the "diffraction-field current"):

$$J_u = J_{PO} + J_d. \tag{6}$$

Since the extension is flat and perfectly conducting, the PO current is known exactly *a priori*. (Note that if the extension is shadowed from the incident field the PO current is simply zero). However, the diffracted field,

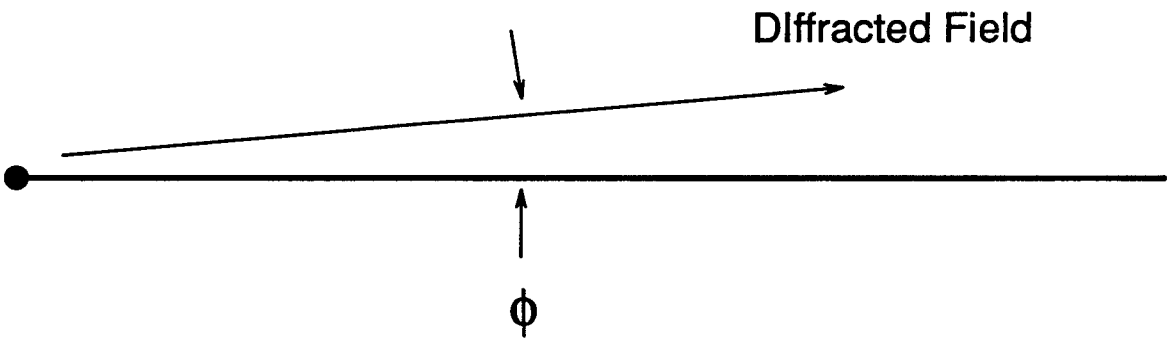


Figure 2: Diffracted field in the vicinity of the extensions.

and therefore the diffraction-field current, is not known initially and must be determined using the moment method. Since it extends to infinity, use of ordinary sub-domain MM basis functions to describe this current would lead to an infinite order system of linear equations that cannot be solved. Instead it is recognized that at distances far enough away from the diffraction point the diffracted field is ray optical. Thus, the form of the diffracted field at the extension beyond points A or D is given by

$$F^d = F_0 \frac{e^{-jkr}}{r} f(\phi), \tag{7}$$

where r is the distance from the diffraction point and $f(\phi)$ is an arbitrary function of the angular cylindrical coordinate with the diffraction point as the origin, as shown in Figure 2. Applying the surface boundary condition $\mathbf{J_s} = \hat{\mathbf{n}} \times \mathbf{H}$ yields the diffraction currents

$$\begin{aligned} J_d &= J_0 \frac{e^{-jkr}}{\sqrt{r}}, \qquad \text{(vertical polarization)} \\ &= J_0 \frac{e^{-jkr}}{r^{1.5}}, \qquad \text{(horizontal polarization).} \end{aligned} \tag{8}$$

We now see that a single basis function of the form of equation (8) can be used to include the diffraction current from the diffraction point to infinity in the hybrid numerical technique. This, combined with the known physical optics currents, entirely describes the current on the infinite extensions. Since there are no discontinuities on the modeled surface, no artificial edge effects are introduced.

The current on the entire surface may now be written as

$$J_u = J_{MM} + J_D + J_{PO}, \tag{9}$$

where J_{MM} is the current between points A and D described by ordinary MM pulse basis functions:

$$J_{MM} = \sum_{m=1}^{N} \alpha_m P(l - l_m), \tag{10}$$

where $P(l - l_m)$ is a pulse function centered at l_m and α_m are unknown weighting coefficients to be found via the moment method. J_D includes both diffraction current terms:

$$J_D = \begin{cases} \alpha_{N+1} J_d, & l < A; \\ \alpha_{N+2} J_d, & l > D. \end{cases} \tag{11}$$

J_{PO} is the physical optics current on the front and back faces given by

$$\mathbf{J_{PO}}(l) = \begin{cases} 2\hat{\mathbf{n}} \times \mathbf{H^i}, & l < A \text{ or } l > D; \\ 0, & \text{elsewhere.} \end{cases} \tag{12}$$

Substituting equation (9) into equation (4) gives

$$-F_z^i(l^-) = L_N^-[J_{MM} + J_D + J_{PO}]. \tag{13}$$

Because the J_{PO} is entirely known *a priori* and $L_N^-[\,]$ is a linear operator, the physical optics term may be moved to the left hand side, giving

$$-F_z^i(l^-) - L_N^-[J_{PO}] = L_N^-[J_{MM}] + L_N^-[J_D]. \tag{14}$$

Thus, the physical optics current simply appears as a field source term in the hybrid technique. Evaluating equation (14) at the centers of the basis functions (point matching or collation), plus at two additional points on the extensions yields $N+2$ algebraically linear equations with $N+2$ unknowns. Solving this system yields the moment weighting coefficients α_m, completing the MM solution of the current. The far field scatter is then determined from

$$F^s = L_N[\rho, J_{MM} + J_D + J_{PO}]\Big|_{\rho \to \infty}, \tag{15}$$

where L_N is L_N^- with the observation point ρ^- replaced with an arbitrary observation point ρ.

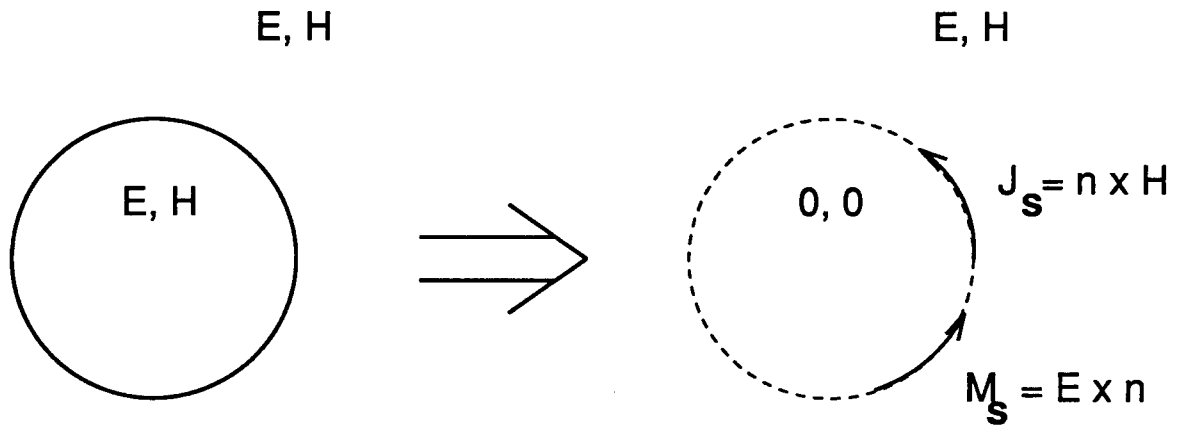


Figure 3: Equivalent problem to be solved with lossy dielectric scatterer.

B. *High Permittivity, High Loss Dielectric Surfaces*

When the scattering surface is perfectly conducting a true surface current exists. Thus, the moment method solves the physical scattering problem directly. When the surface is not perfectly conducting a surface current cannot be supported; the field penetrates the surface and a volume current density exists. The moment method is not well suited for direct application to volume current problems. Instead, the equivalence principle [18] is applied as shown in Figure 3, yielding both electric (J) and magnetic (M) surface current densities that radiate the desired scattered field. Assuming that the conditions

$$|N| \gg 1, \qquad |\mathrm{Im}(N) k \rho_l| \gg 1 \tag{16}$$

where N is the complex refractive index of the scattering medium and ρ_l is the radius of curvature of the surface, are met everywhere on the surface, the electric and magnetic current densities can be related by impedance boundary conditions [19]. Because of the high refractive index, the field penetrating into the surface propagates as a plane wave in the negative surface normal direction. The two surface current components can then be related by [20]

$$\mathbf{M} = -Z_s \hat{\mathbf{n}} \times \mathbf{J}, \tag{17}$$

where Z_s is the intrinsic wave impedance of the lossy dielectric.

A MFIE may be derived for the impedance boundary by setting the field radiated by both the electric and magnetic surface currents equal to the negative of the incident field just below the interface. For a

one-dimensional interface and vertical polarization, this gives

$$-H_z^i(l^-) = H_z^s[\rho^-, J_l(l)] + H_z^s[\rho^-, M_z(l)]. \tag{18}$$

The first term on the right hand side is given by (1). The second term can be simplified by applying duality to equation (3) and using (17) to give [21]

$$-H_z^i(l) = L_1^-[J_l(l)] - \frac{Z_s}{\eta^2} L_2^-[J_l(l)]. \tag{19}$$

Similarly, with a lossy dielectric surface the EFIE becomes

$$-E_z^i(l^-) = L_2^-[J_z(l)] + Z_s L_1^-[J_z(l)]. \tag{20}$$

Equations (19) and (20) can be solved for J_u using moment method techniques.

The hybrid MM/GTD technique can be extended to apply to equations (19) and (20) to find the scattering from lossy dielectric surfaces of the type shown in Figure 1 with little modification. The surface current between points A and D is again divided into pulse basis functions as described in equation (10), and the diffraction-current basis functions are unchanged from equation (8) since the diffracted field is still ray optical at suitable distances from the diffraction point [22]. The physical optics current does need to be modified slightly since the surface is no longer perfectly conducting:

$$\mathbf{J_{PO}}(l) = \begin{cases} (1-\Gamma)\,\hat{\mathbf{n}} \times \mathbf{H^i}, & l < A \text{ or } l > D; \\ 0, & \text{elsewhere,} \end{cases} \tag{21}$$

where Γ is the appropriate parallel (vertical) or perpendicular (horizontal) polarized reflection coefficient on the front and back extensions. (Note that equation (21) reduces to (12) with a perfectly conducting surface.) Substituting equation (9) (with the modified J_{PO}) into equation (19) and moving the known terms to the left hand (source) side yields

$$-H_z^i(l^-) - L_1^-[J_{PO}(l)] + \frac{Z_s}{\eta^2} L_2^-[J_{PO}(l)] \tag{22}$$

$$= L_1^-[J_{MM}(l) + J_D(l)] - \frac{Z_s}{\eta^2} L_2^-[J_{MM}(l) - J_D(l)]. \tag{23}$$

Similarly, the EFIE becomes

$$E_z^i(l^-) - L_2^-[J_{PO}(l)] - Z_s L_1^-[J_{PO}(l)] \tag{24}$$

$$= L_2^-[J_{MM}(l) + J_D(l)] + Z_s L_1^-[J_{MM}(l) - J_D(l)]. \tag{25}$$

Both equations (22) and (24) can be evaluated at the $N+2$ matching points, and the resulting linear system algebraic equations solved to give the unknown coefficients α_n, completing the numerical solution. The far-field scattering from the surface is then found by evaluating

$$H^s = L_1[\rho, J_{MM} + J_D + J_{PO}]\Big|_{\rho\to\infty} + Z_s L_2[\rho, J_{MM} + J_D + J_{PO}]\Big|_{\rho\to\infty} \tag{26}$$

or

$$E^s = L_2[\rho, J_{MM} + J_D + J_{PO}]\Big|_{\rho\to\infty} - \frac{Z_s}{\eta^2} L_1[\rho, J_{MM} + J_D + J_{PO}]\Big|_{\rho\to\infty} \tag{27}$$

C. *Low Permittivity, Low Loss Dielectric Surfaces*

When the conditions of equation (16) are not met, impedance boundary conditions can not be used to represent the interface. Instead, a numerical implementation that is valid for a general dielectric interface must be used. Our approach is based on that of [23]. The scattering problem, shown in the left illustration of Figure 4, can be separated into external and internal equivalents, as shown in the middle and right illustration, respectively. External sources $\mathbf{J^i}$ and $\mathbf{M^i}$ in the original problem radiate in free space in the presence of an obstacle with electric parameters ϵ_1 and μ_1. The external fields are the sum of the incident and scattered fields,

$$\begin{aligned} \mathbf{E} &= \mathbf{E^i} + \mathbf{E^s} \\ \mathbf{H} &= \mathbf{H^i} + \mathbf{H^s}, \end{aligned} \tag{28}$$

and the internal fields are $\mathbf{E^1}$ and $\mathbf{H^1}$.

The external equivalent model is used to calculate the fields external to the dielectric boundary. Here, the obstacle is removed and replaced by surface current densities

$$\mathbf{J} = \hat{\mathbf{n}} \times \mathbf{H},$$

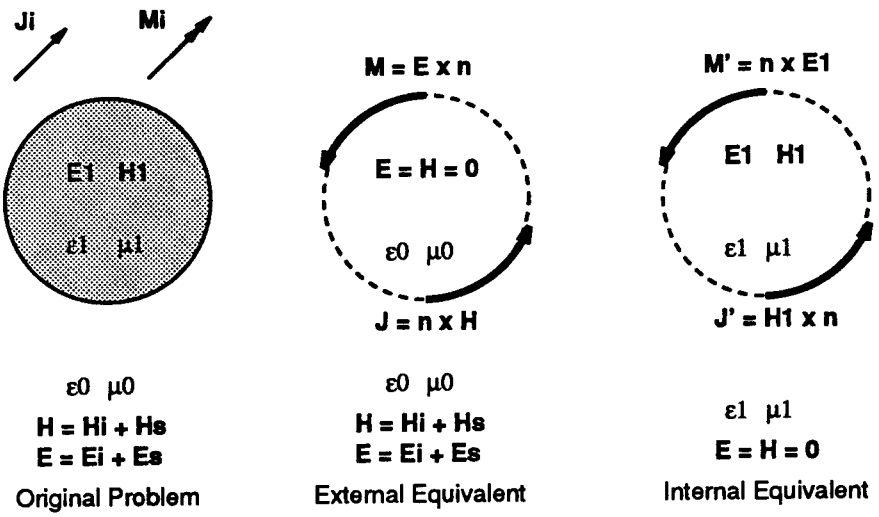


Figure 4: External and internal equivalent problems.

$$\mathbf{M} = \mathbf{E} \times \hat{\mathbf{n}}, \tag{29}$$

placed along the obstacle boundary. The internal electric and magnetic fields are chosen to be zero, and the total fields external to the obstacle boundary $\mathbf{E}$ and $\mathbf{H}$ are chosen to be the external fields of the original problem.

Likewise, the internal equivalent model is used to determine the fields transmitted into the dielectric region. In this case, the external space is replaced by the dielectric material, and surface current densities

$$\begin{aligned} \mathbf{J}' &= \mathbf{H}^1 \times \hat{\mathbf{n}}, \\ \mathbf{M}' &= \hat{\mathbf{n}} \times \mathbf{E}^1, \end{aligned} \tag{30}$$

are placed on the boundary of the dielectric obstacle. The exterior electric and magnetic fields are now chosen to be zero, and the interior fields are chosen to be the interior fields of the original problem, $\mathbf{E}^1$ and $\mathbf{H}^1$.

The boundary conditions on the tangential electric and magnetic fields,

$$\begin{aligned} \hat{\mathbf{n}} \times \mathbf{E} - \hat{\mathbf{n}} \times \mathbf{E}^1 &= 0, \\ \hat{\mathbf{n}} \times \mathbf{H} - \hat{\mathbf{n}} \times \mathbf{H}^1 &= 0, \end{aligned} \tag{31}$$

must be satisfied in order to validate the external and internal equivalent models. Substituting $-\mathbf{M} = \hat{\mathbf{n}} \times \mathbf{E}$,

$\mathbf{M}' = \hat{\mathbf{n}} \times \mathbf{E}^1$, $\mathbf{J} = \hat{\mathbf{n}} \times \mathbf{H}$, and $-\mathbf{J}' = \hat{\mathbf{n}} \times \mathbf{H}^1$ into equation (31) leads to $\mathbf{M}' = -\mathbf{M}$ and $\mathbf{J}' = -\mathbf{J}$.

The integral equations are formed using both the internal and external equivalent models. Just inside the boundary in the external equivalent model, the scattered field radiated by $\mathbf{J}$ and $\mathbf{M}$ must cancel out the incident field, expressed as

$$-\mathbf{F}^{\mathbf{i}} = \mathbf{F}^{\mathbf{s}}(\mathbf{J}, \mathbf{M})|_{S^-}^{ext}, \tag{32}$$

where F represents the field component (either E or H), *tan* indicates the tangential components, S^- indicates a surface just inside the boundary, and *ext* refers to the use of the *external* electrical parameters. In the internal equivalent model, the scattered fields radiated by $\mathbf{J}' = -\mathbf{J}$ and $\mathbf{M}' = -\mathbf{M}$ must be zero outside of the boundary. This relationship is represented by

$$\mathbf{F}^{\mathbf{s}}(\mathbf{J}', \mathbf{M}')|_{tan,S^+}^{int} = \mathbf{F}^{\mathbf{s}}(-\mathbf{J}, -\mathbf{M})|_{tan,S^+}^{int} = 0, \tag{33}$$

where S^+ indicates a surface just outside the boundary and *int* refers to the use of the *internal* electrical parameters.

For the one-dimensionally rough interfaces of interest here, with vertically polarized illumination equations (32) and (33) reduce to

$$\begin{aligned} -H_z^i &= L_2^{-\prime}[M_z] + L_1^-[J_l] \\ 0 &= L_2^{+\prime}[M_z] + L_1^+[J_l], \end{aligned} \tag{34}$$

respectively. The superscript $^+$ indicates that the operator L_N is evaluated immediately *above* the interface using the dielectric properties of the internal medium. (As before, $^-$ indicates that the evaluation is just *below* the interface with the dielectric properties of the external medium used.) A prime indicates that the dual of the operator is actually used. Similarly, horizontally polarized illumination yields

$$\begin{aligned} -E_z^i &= L_2^-[J_z] - L_1^{-\prime}[M_l], \\ 0 &= L_2^+[J_z] - L_1^{+\prime}[M_l]. \end{aligned} \tag{35}$$

Equations (34) and (35) can be solved using the MM/GTD approach. Again, the surface currents are

divided into moment method, physical optics, and diffraction current components:

$$\begin{aligned} J_u &= J_{MM} + J_D + J_{PO}, \\ M_u &= M_{MM} + M_d + M_{PO}. \end{aligned} \tag{36}$$

The forms of J_{MM}, J_D, and J_{PO} are unchanged from that used for the high loss surface with impedance boundary conditions. The moment method magnetic current is divided into $N+2$ basis functions as

$$M_{MM} = \sum_{m=1}^{N} \alpha_{m+N+2} P(l - l_m), \tag{37}$$

and the physical optics magnetic current is

$$\mathbf{M_{PO}}(l) = \begin{cases} (1+\Gamma)\,\mathbf{E^i} \times \hat{\mathbf{n}}, & l < A \text{ or } l > D; \\ 0, & \text{elsewhere.} \end{cases} \tag{38}$$

The magnetic diffraction current is given by

$$M_D = \begin{cases} \alpha_{2N+3} M_d, & l < A; \\ \alpha_{2N+4} M_d, & l > D; \end{cases} \tag{39}$$

where

$$\begin{aligned} M_d &= \frac{e^{-jk_0 r}}{r^{1.5}}, \qquad \text{(vertical polarization)}, \\ &= \frac{e^{-jk_0 r}}{\sqrt{r}}, \qquad \text{(horizontal polarization)}, \end{aligned} \tag{40}$$

where k_0 is the wave number above the surface.

Note that since the wave numbers above and below the surface are different, the diffraction currents of equations (8) and (40) do not meet the surface boundary conditions for the diffracted field exactly. However, use of different forms of the diffraction basis functions had only small effects on the calculated scattering (typically less than 1 dB), indicating that the basis function is needed primarily to eliminated the discontinuity in the current, but otherwise does not contribute significantly to the far-field scatter. Also note that this general dielectric implementation of the MM/GTD technique leads to twice as many unknowns ($2N+4$) as the impedance boundary and perfectly conducting implementations. Thus, although the general implementation

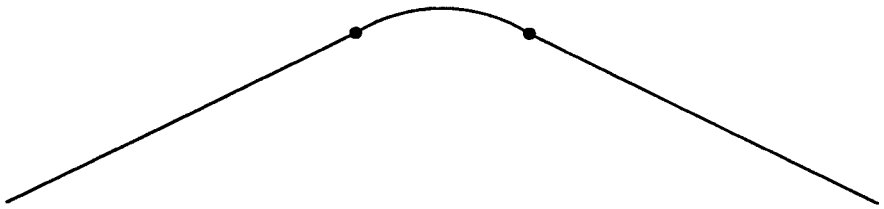


Figure 5: Rounded wedge.

can be applied to any type surface, dramatic computational savings are realized in using the impedance boundary approach, if possible.

1) *Example application:* Efficient numerical approaches in implementing the hybrid MM/GTD for the impedance boundary conditions were discussed in [17]. These are also applicable to the general dielectric implementation. The validity of the perfectly conducting and impedance boundary implementations of the hybrid MM/GTD technique were confirmed by direct comparison with the GTD-predicted diffraction from wedges in [10] and [24]. We therefore only concern ourselves with the low-loss dielectric implementation here.

There are no problems of the type shown in Figure 1 for which a solution (exact or asymptotic) is known to the investigators when the surface represents an arbitrary dielectric interface. Thus, we compare the predictions of the general dielectric implementation with those of the impedance boundary under conditions where both are expected to be valid. The scattering from a rounded dielectric wedge of the type shown in Figure 5 is used. The internal angle of the wedge is 120° and the radius of curvature of the apex is 1 λ. Figure 6 shows the backscattering calculated when the complex wedge permittivity is $30 - j1$. (The grazing illumination angle is relative to horizontal). Very good agreement is achieved between the two approaches. The maximum difference is only 1.5 dB, and occurs at horizontal polarization only when the backscattering cross-section is quite small. This discrepancy likely occurs because the the first condition of equation (16) is only marginally met with these dielectric parameters, so the impedance boundary results are slightly in error. No evidence of edge effects appear in any scattering curve.

III. APPLICATION

The scattering calculations were applied to the same surface configurations that were used by West [10]. Two different surfaces are used. The first is shown in Figure 7. The range from point F to point H is based one cycle of a near-breaking ("Stokes") ocean wave, defined using the approximate function of Longuet-Higgins

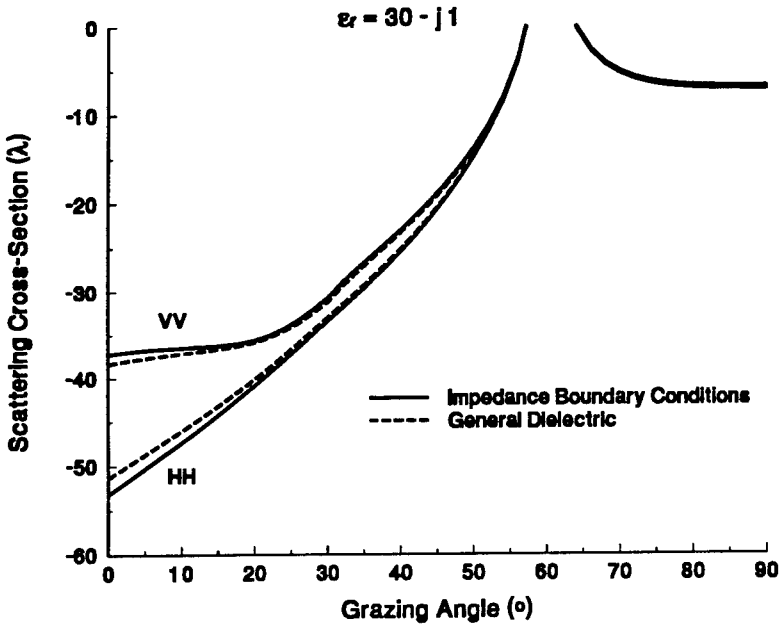


Figure 6: Scattering from rounded wedge when $\epsilon_r = 30 - j1$.

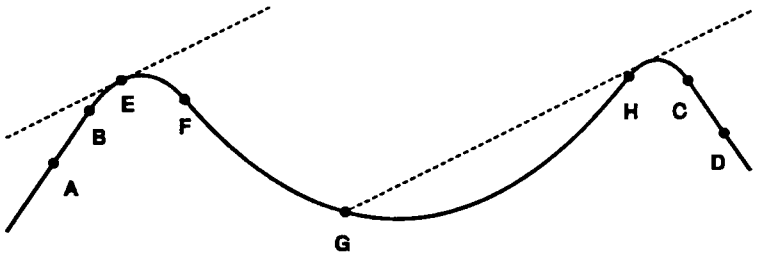


Figure 7: Weakly shadowing surface. The dashed lines show the shadow boundaries with 10° grazing illumination.

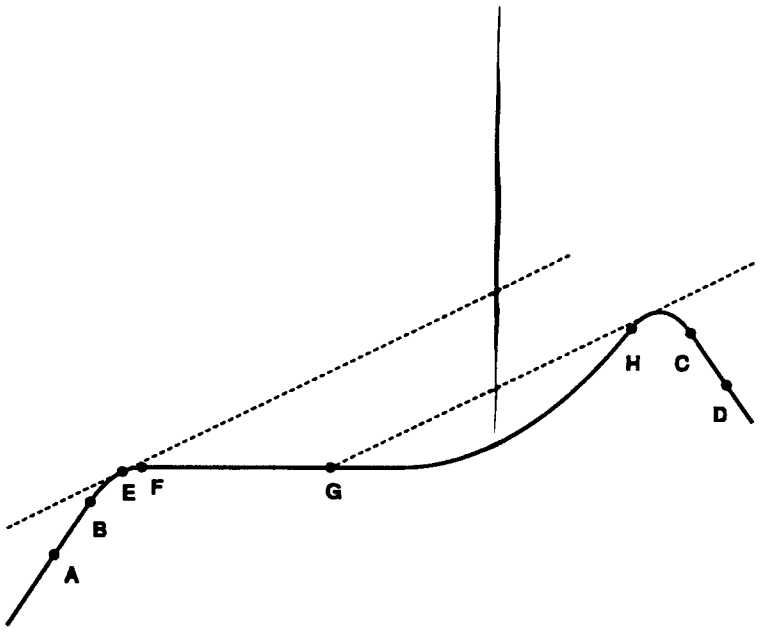


Figure 8: Deeply shadowing surface. The dashed lines show the shadow boundaries with 10° grazing illumination.

[25]:

$$y' = ln(\sec x'), \qquad |x'| \leq \pi/6, \tag{41}$$

where x' and y' are normalized coordinates of the Stokes wave. The surface is then rounded off from points H to C (B to F), and extended to infinity beyond point D (A). The extensions angle downward at 30° to horizontal. The calculations were performed with the distance between the two crests (the "Stokes wavelength") set at 25 λ, where λ is the electromagnetic wavelength. The radius of curvature from point B to E was set at 5 λ, and the radius from point H to C set at 2 λ, where λ is the electromagnetic wavelength. The back crest of the surface is only a few tenths of a wavelength below the shadow boundary when the illumination is at near grazing incidence, so this surface is termed the "weakly shadowing surface". The second surface was formed by setting the displacement of the back half of the Stokes wave to zero, as shown in Figure 8. The radii of curvature between points B and E and F and C are again 5 λ and 2 λ respectively.

As in [10], a random small-scale roughness was added to the large-scale surfaces of Figures 7 and 8 to give a Bragg-resonant backscatter. Two configurations were used. In the "rough-in-shadow" case the small-scale roughness was extended from point B to point H, while in the smooth-in-shadow case the roughness extended only from points E to G. (The roughness was actually extended slightly into the shadowed regions to avoid an

unrealistic discontinuity at that point. The standard deviation of the roughness was tapered smoothly to zero at a distance of 0.5 λ within the shadowed region.) The small-scale displacement was numerically generated from a Gaussian power spectral density that had a correlation length of 0.2 λ and a height standard deviation of 0.045 λ. The scattering from 40 independent realizations of the small-scale roughness was averaged to remove phase-interference fading.

The numerical technique was applied using surface dielectric properties chosen to approximate soils under different conditions. Three complex dielectric constants were used: $3 - j0$ to represent dry sand, $10 - j2$ for "typical soil", and $35 - j5$ for wet soils [26]. The calculations were also repeated with a perfectly conducting surface to provide a benchmark for comparison. Pulse basis functions of length 0.05 λ were used to describe the surface currents in the moment method section of the surface (between points A and D). Point A (D) was 0.5 λ from point B (C). The scattering predicted by the two-scale scattering model was predicted by integrating the small-perturbation scattering coefficients of Ulaby *et al.* [27] over the arc length of the directly illuminated surface. The local incidence angle was adjusted by the large-scale tilt of the surface.

IV. RESULTS

A. *Backscattering*

1) *Weak-shadowing surface:* The scattering from the perfectly conducting weak-shadowing surface is shown in Figure 9. At vertical polarization the rough-in-shadow and smooth-in-shadow scattering show good agreement down to 7° grazing. At smaller grazing the smooth-in-shadow scattering drops much more rapidly, indicating that the shadow region roughness significantly contributes to the scattering here. At horizontal polarization the agreement extends down to 2°, where the smooth-in-shadow scattering drops more quickly. At both polarizations the two-scale model loses accuracy below 10° grazing, showing that this approach is quite limited at small grazing even when corrected for shadowing. Reasons for the loss of accuracy are discussed in [10].

The results when the scattering medium has relative permittivities of $35 - j5$, $10 - j2$, and $3 - j0$ are shown in Figures 10 through 12. As the magnitude of the dielectric constant drops the overall magnitude of the backscatter also drops as expected, and the levels of vertically and horizontally polarized backscattering become closer. Also, the rough-in-shadow and smooth-in-shadow agree down to 3° at vertical polarization in all cases. The agreement at horizontal polarization is not significantly different than that observed with the

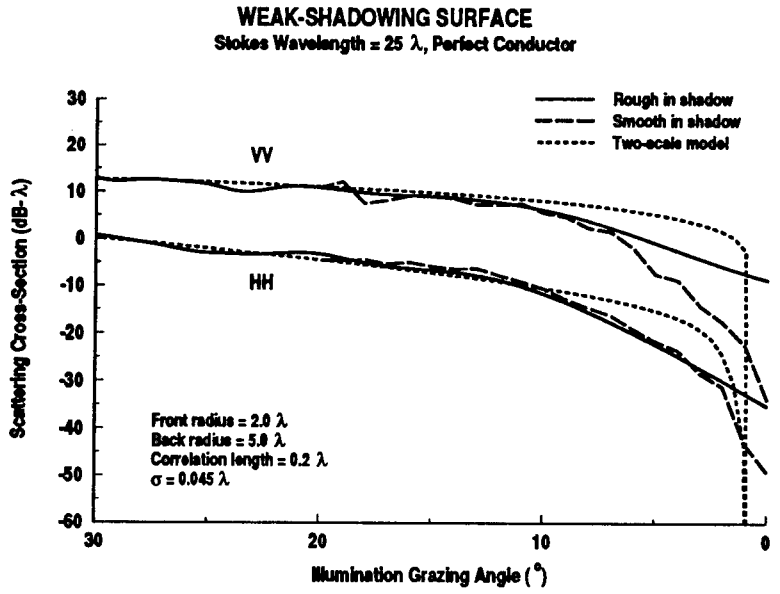


Figure 9: Backscattering from perfectly conducting weak-shadowing surface.

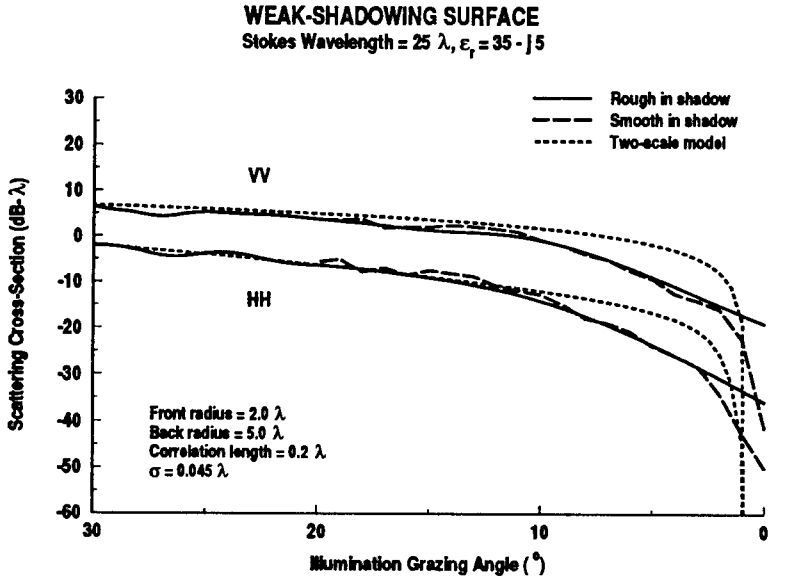


Figure 10: Backscattering from weakly shadowing "moist clay" surface.

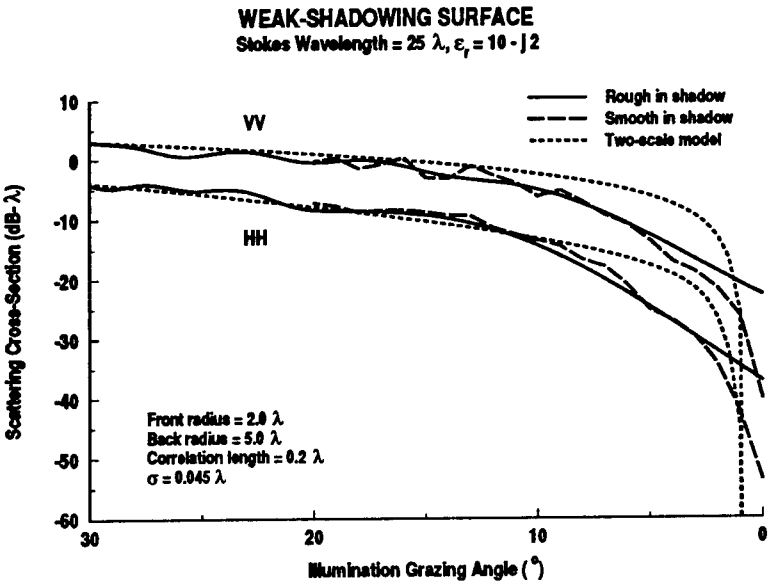


Figure 11: Backscattering from weakly shadowing "typical soil" surface.

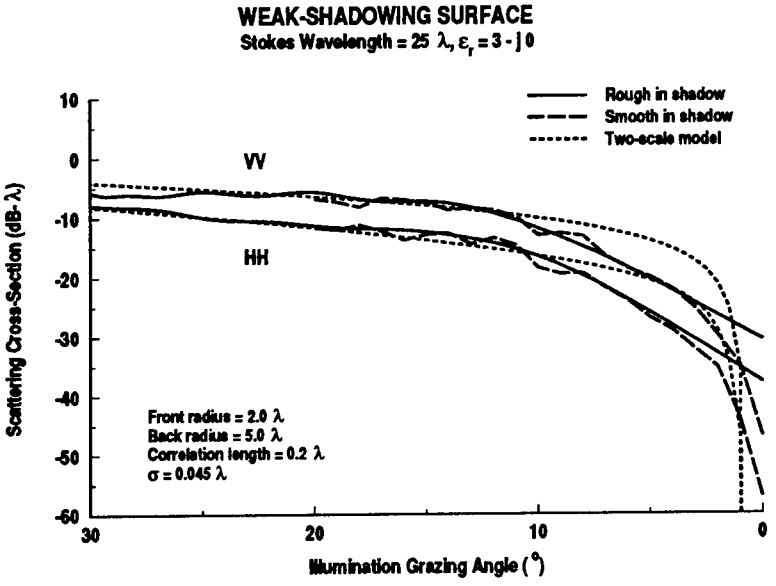


Figure 12: Backscattering from weakly shadowing "dry sand" surface.

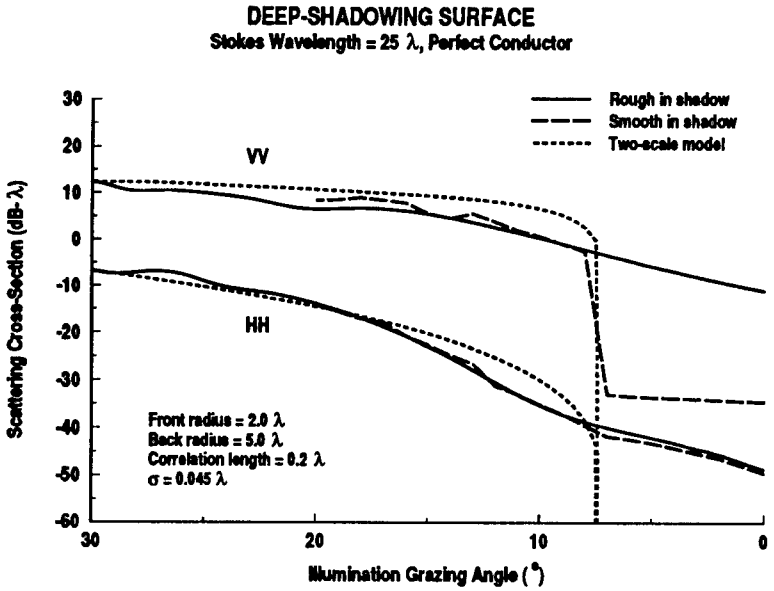


Figure 13: Backscattering from perfectly conducting deep-shadowing surface.

perfectly conducting surface. No change is observed in the accuracy of the two-scale model with the changing permittivity at either polarization.

2) *Deep-shadowing surface:* The scattering from the perfectly conducting deep-shadowing surface is shown in Figure 13. Here, the rough-in-shadow and smooth-in-shadow scattering shows close agreement at vertical polarization down to 8°, where the smooth-in-shadow scattering drops abruptly. Beyond this angle the shadow boundary from the front crest no longer intersects the back part of the surface, so small-scale roughness is no longer included in the smooth-in-shadow surface. Also, all small-scale roughness on the rough-in-shadow surface is shadowed. The large differences between the smooth-in-shadow and rough-in-shadow scattering here show that deeply-shadow roughness contributes significantly at vertical polarization. At horizontal polarization, the rough-in-shadow and smooth-in-shadow scattering shows good agreement (to within 2 dB) at all grazing angles, indicating that deeply-shadowed roughness can be ignored at this polarization.

The scattering calculated using the dielectric surfaces is shown in Figures 14 through 16. At vertical polarization, the agreement between the rough-in-shadow and smooth-in-shadow scattering at all dielectric constants is similar to that calculated with the perfectly conducting surface. The agreement is good until

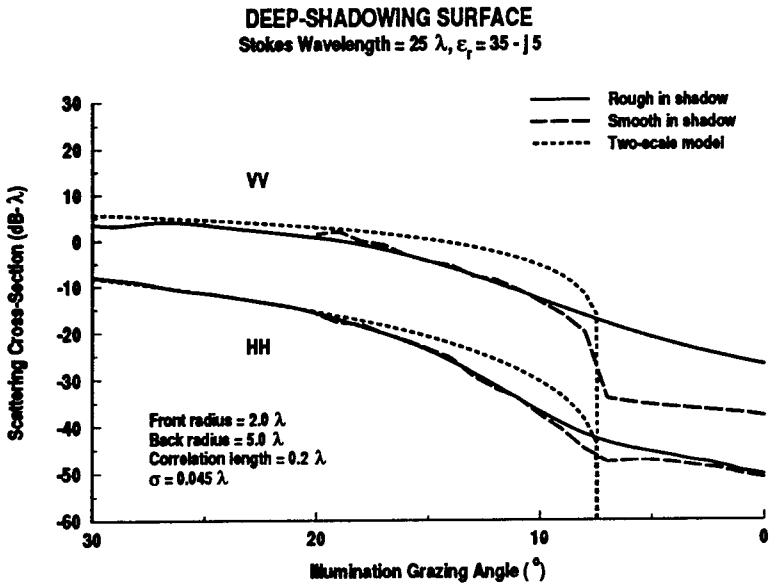


Figure 14: Backscattering from deeply shadowing "moist clay" surface.

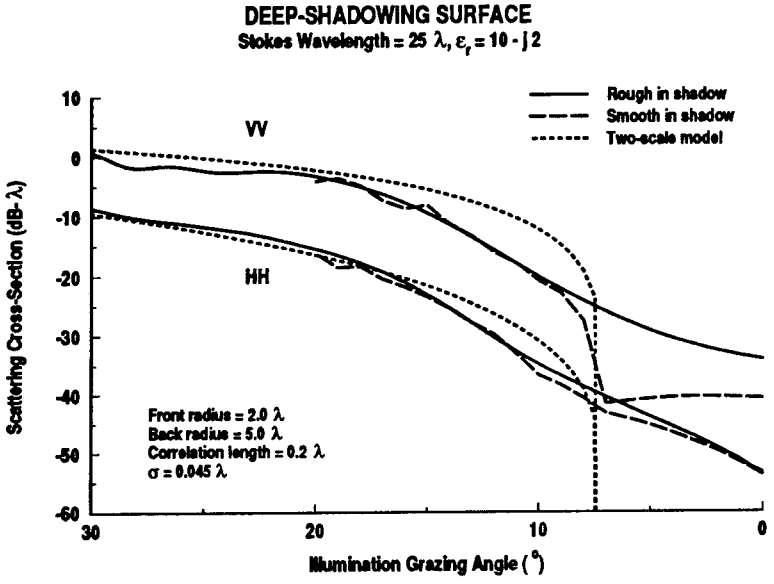


Figure 15: Backscattering from deeply shadowing "typical soil" surface.

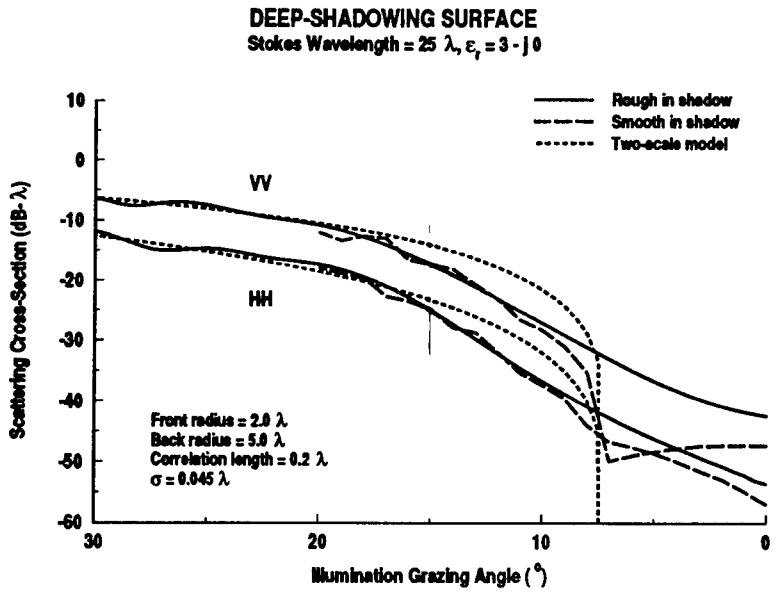


Figure 16: Backscattering from deeply shadowing "dry sand" surface.

just above 8° grazing, where the smooth-in-shadow scattering drops dramatically. The rough-in-shadow scattering continues to drop smoothly beyond 8°. However, as the magnitude of the dielectric constant drops, the scattering within this entirely-shadowed region drops more rapidly with decreasing grazing. This suggests that the shadow-region roughness is less important at vertical polarization with decreasing (complex) permittivity.

At horizontal polarization, the rough-in-shadow and smooth-in-shadow scattering agrees to 8° grazing at all permittivities examined. At the higher permittivities ($35 - j5$ and $10 - j2$), this agreement is maintained to within 2 dB down to 0°. At $3 - j0$ permittivity, however, the rough-in-shadow scattering is slightly greater than the smooth-in-shadow, indicating that the deeply shadowed roughness is more important with this dielectric constant than it is with a perfectly conducting surface.

B. *Surface Currents*

The backscattering calculations suggest that the shadowed-region roughness becomes less important with decreasing dielectric constant at vertical, but more important at horizontal polarization. To further examine this, the surface electric currents calculated with the deep-shadowing surfaces are shown in Figures 17 through 20. The illumination grazing angle is 0° in each case, and the current levels are normalized to the current

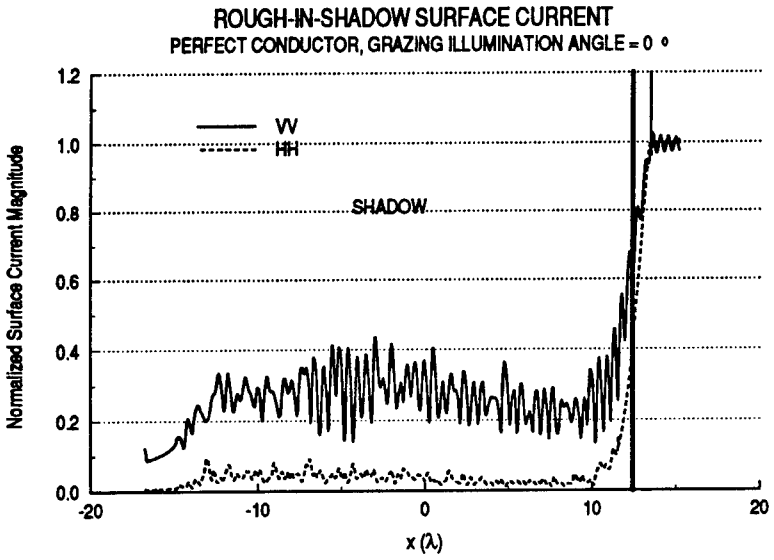


Figure 17: Current magnitudes for $\theta_g = 0°$, perfectly conducting deep-shadowing surface.

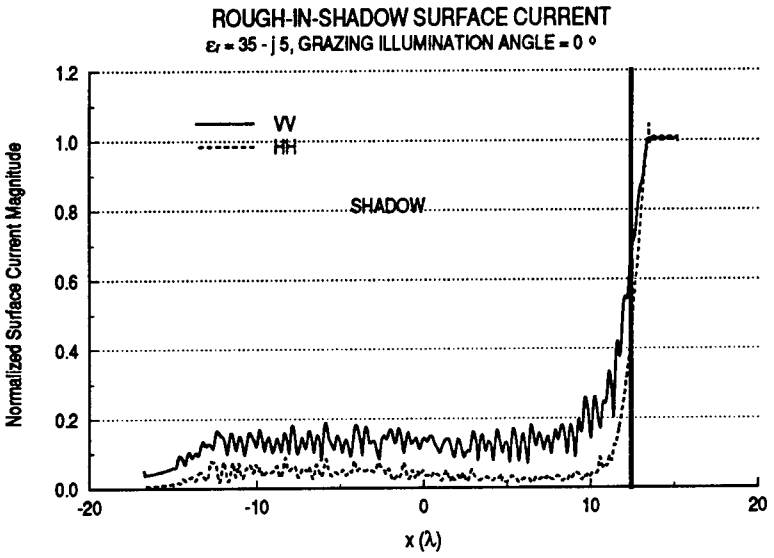


Figure 18: Current magnitudes for $\theta_g = 0°$, $\epsilon_r = 35 - j5$, deep-shadowing surface.

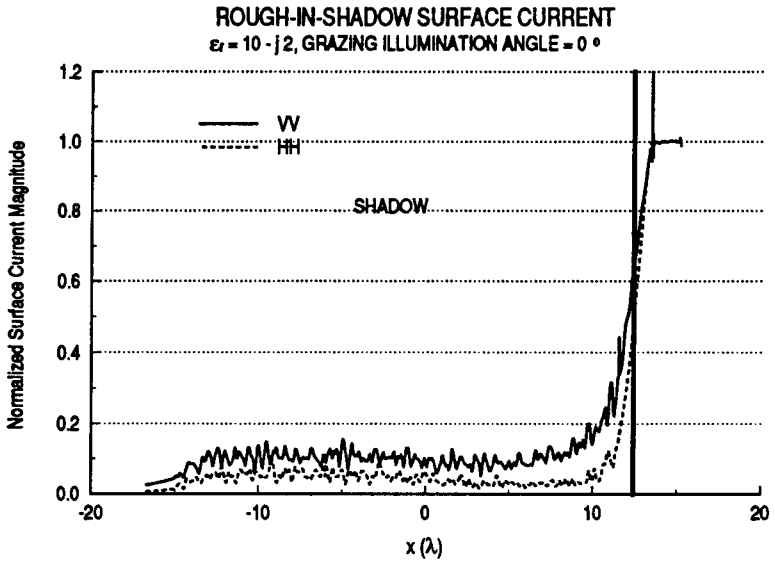


Figure 19: Current magnitudes for $\theta_g = 0°$, $\epsilon_r = 10 - j2$, deep-shadowing surface.

given by the physical optics approximation on the directly illuminated (front) infinite surface extension. With the perfectly conducting surface, the currents associated with the vertically polarized illumination are much higher than those associated with horizontal illumination in the shadow region as expected. When $\epsilon_r = 35 - j5$, the vertical polarization currents drop dramatically, while the horizontal polarization currents increase slightly. This trend continues until, at $\epsilon_r = 3 - j0$, the vertical and horizontal polarized currents have nearly equal magnitudes.

V. SUMMARY AND CONCLUSIONS

A numerical study has been performed to investigate the validity of shadowing functions based on geometrical optics shadowing in predicting the electromagnetic scattering from rough interfaces of differing dielectric properties. Two previously developed implementations of a hybrid moment method/geometrical theory of diffraction numerical technique were used, and a new version was implemented that allowed the treatment of scattering media with arbitrary (homogeneous) dielectric properties. The effect of both weakly and strongly shadowed small-scale roughness on the backscatter was examined.

The numerical calculations showed that weakly-shadowed roughness can contribute significantly to the

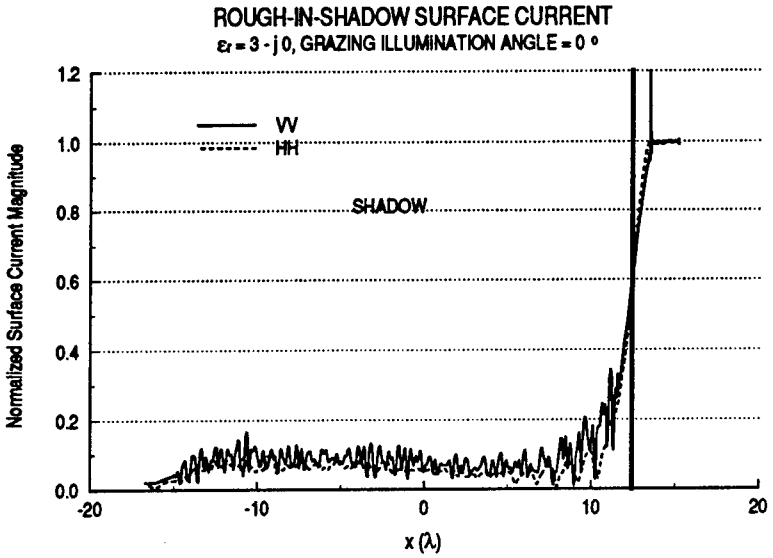


Figure 20: Current magnitudes for $\theta_g = 0°$, $\epsilon_r = 3 - j0$, deep-shadowing surface.

total scattering at both vertical and horizontal polarization. The contribution appears to be strongest at vertical polarization with a perfectly conducting interface, and decreases as the magnitude of the surface permittivity drops. At horizontal polarization the relative importance of weakly shadowed roughness does not depend strongly on the surface dielectric properties. Deeply shadowed roughness can contribute significantly to the backscattering from surfaces at vertical polarization. This effect is most prominent with perfectly conducting surfaces, and becomes less important as the surface permittivity drops. Deeply shadowed roughness is less important at horizontal polarization, and can be ignored when the surface permittivity is large (or the surface is perfectly conducting). Deeply shadowed roughness did raise the horizontally polarized scattering slightly when the surface dielectric constant was reduced to $3 - j0$, but the effect was much less dramatic than that observed at vertical polarization at the same incidence. As was earlier demonstrated with perfectly conducting surfaces, use of a shadowing function offers very little, if any, improvement in the accuracy of the two-scale model when surface-self shadowing occurs, independent of the surface dielectric properties or illumination polarization.

REFERENCES

[1] P. Beckmann and A. Spizzichino, *The Scattering of Electromagnetic Waves from Rough Surfaces*, Pergamon, New York, 1963.

[2] S. O. Rice, "Reflection of electromagnetic wave from slightly rough surfaces", *Communications in Pure and Applied Mathematics*, vol. 4, no. 2, pp. 351–378, Aug. 1951.

[3] G. R. Valenzuela, "Scattering of electromagnetic waves from a tilted, slightly rough surface", *Radio Science*, vol. 3, no. 11, pp. 1057–1066, Nov. 1968.

[4] P. Beckmann, "Shadowing of random rough surfaces", *IEEE Transactions on Antennas and Propagation*, vol. AP-13, no. 3, pp. 384–388, May 1965.

[5] R. A. Brockelman and T. Hagfors, "Note on the effect of shadowing on the backscattering of waves from a random rough surface", *IEEE Transactions on Antennas and Propagation*, vol. AP-14, no. 5, pp. 621–629, 1966.

[6] M. L. Sancer, "Shadow-corrected electromagnetic scattering from a randomly rough surface", *IEEE Transactions on Antennas and Propagation*, vol. AP-17, no. 5, pp. 577–585, Sept. 1969.

[7] L. B. Wetzel, "Electromagnetic scattering from the sea at low grazing angles", in *Surface Waves and Fluxes*, G. L. Geernaert and W. L. Plant, Eds., vol. II—Remote Sensing, pp. 109–171. Kluwer, Dordrecht, The Netherlands, 1990.

[8] D. E. Barrick, "Near-grazing illumination and shadowing of rough surfaces", *Radio Science*, vol. 30, no. 3, pp. 563–580, May 1995.

[9] D. Holliday, L. L. DeRaad, and G. J. St. Cyr, "Volterra approximation for low grazing angle shadowing on smooth ocean-like surfaces", *IEEE Transactions on Antennas and Propagation*, vol. 43, no. 9, pp. 1199–1206, Nov. 1995.

[10] J. C. West, "Effect of shadowing on electromagnetic scattering from rough ocean-wave-like surface at small grazing angles", *IEEE Transactions on Geoscience and Remote Sensing*, 1997, in press.

[11] L. M. Mockapetris, "Effect of surface and scattering parameters on two scale of roughness models", in *Proceeding of the Progress in Electromagnetic Research Symposium*, July 24–28, Seattle, Washington, 1995, p. 224.

[12] A. K. Fung and M. F. Chen, "Numerical simulation of scattering from simple and composite random surfaces", *Journal of the Optical Society of America, Series A*, vol. 2, no. 12, pp. 2274–2284, Dec. 1985.

[13] M. F. Chen and A. K. Fung, "A numerical study of the regions of validity of the Kirchhoff and small-perturbation rough surface scattering models", *Radio Science*, vol. 23, no. 2, pp. 163–170, Mar. 1988.

[14] E. I. Thorsos, "The validity of the Kirchhoff approximation for rough surface scattering using a Gaussian roughness spectrum", *Journal of the Acoustical Society of America*, vol. 83, no. 1, pp. 78–82, Jan. 1988.

[15] Y. Kim, E. Rodriguez, and S. Durden, "A numerical assessment of rough surface scattering theories: Vertical polarization", *Radio Science*, vol. 27, no. 4, pp. 515–527, July 1992.

[16] W. D. Burnside, C. L. Yu, and R. J. Marhefka, "A technique to combine the geometrical theory of diffraction and the moment method", *IEEE Transactions on Antennas and Propagation*, vol. AP-23, no. 4, pp. 551–558, July 1975.

[17] J. C. West and J. M. Sturm, "A hybrid MM/GTD numerical technique for lossy dielectric rough surface scattering calculations", Final Report for AFOSR Summer Faculty Research Program, Rome Laboratory, Hanscom AFB, 1995.

[18] C. A. Balanis, *Advanced Engineering Electromagnetics*, Wiley, New York, 1989.

[19] T. B. A. Senior and J. L. Volakis, "Generalized impedance boundary conditions in scattering", *Proceedings of the IEEE*, vol. 79, no. 10, pp. 1413–1420, Oct. 1991.

[20] A. W. Glisson, "Electromagnetic scattering by arbitrary shaped surfaces with impedance boundary conditions", *Radio Science*, vol. 27, no. 6, Nov. 1992.

[21] W. V. T. Rusch and R. P. Pogorzelski, "A mixed-field solution for scattering from composite bodies", *IEEE Transactions on Antennas and Propagation*, vol. AP-34, no. 7, 1986.

[22] R. Tiberio, G. Pelosi, G. Manara, and P. H. Pathak, "High-frequnency scattering from a wedge with impedance faces illuminated by a line source, part i: Diffraction", *IEEE Transactions on Antennas and Propagation*, vol. 37, no. 2, Feb. 1989.

[23] E. Arvas, S. M. Rao, and T. K. Sarkar, "E-field solution of tm-scattering from multiple perfectly conducting and lossy dielectric cylinders of arbitrary cross-section", *IEE Proceedings*, vol. 133, Pt. H, no. 2, pp. 115–121, Apr. 1986.

[24] J. C. West and J. M. Sturm, "A hybrid mm/gtd numerical technique for far-field scattering from impedance boundaries", in *IEEE Antennas and Propagation Society International Symposium*, July 21–26, Hyatt Regency Hotel, Baltimore, MD, 1996.

[25] M. S. Longuet-Higgins, "On the form of the highest progressive and standing waves in deep water", *Proceeding of the Royal Society of London, Series A*, vol. 331, no. 1587, pp. 445–456, 1973.

[26] F. T. Ulaby, R. K. Moore, and A. K. Fung, *Microwave Remote Sensing: Active and Passive*, vol. 3, Artech House, Norwood, Massachusetts, 1986.

[27] F. T. Ulaby, R. K. Moore, and A. K. Fung, *Microwave Remote Sensing: Active and Passive*, vol. 2, Artech House, Norwood, Massachusetts, 1982.